Diffusion approximation as a tool in computer networks performance evaluation

CNT 2500

Tadeusz Czachórski

tadek@iitis.pl

https://www.iitis.pl/en/person/tczachorski

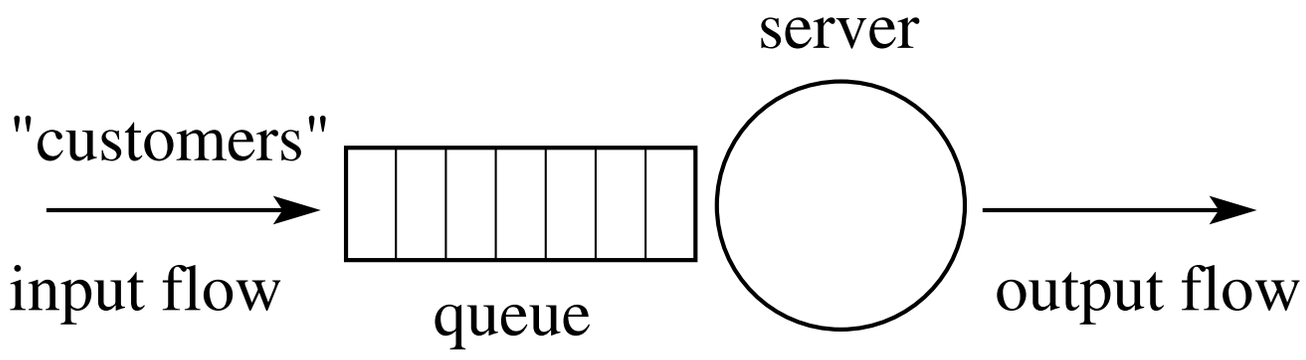
Institute of Theoretical and Applied Informatics
of Polish Academy of Sciences, Gliwice, Poland

The 9th International Conference on Electronics, Communications and Networks
CECNet 2019, October 18-21, 2019 Kitakyushu City, Japan

Introduction

- Since Erlang and Engset times, queueing models are extensively used to model telephone, computer and telecommunication networks.
- The rapid growth of telephone switching systems in offices of public switch telephone networks (PSTN), mobile switching centers (MSC) and radio access networks of cellular mobile networks, cloud data center systems and customer service centers has renewed the interest in multiserver queueing models
- Markov chain models consider events thus have limitations (explosion of the number of states), we propose *steady state and transient state diffusion approximation models* based on changes of flows, giving numerical results.
- We concentrate on *the G/G/c/c+K queueing model* (Kendal's notation) having in mind *design and performance evaluation* of multichannel communication networks.

Queueing model, generic problem



Known:
– arrival pattern, e.g. interarrival time distribution
– service time, e.g. service time distribution
– queueing discipline
– queue size limitations

To determine:
– queue distribution (or its moments)
– waiting time distribution (or its moments)
– losses (if limited queue)

Some Applications of G/G/c/c+K models

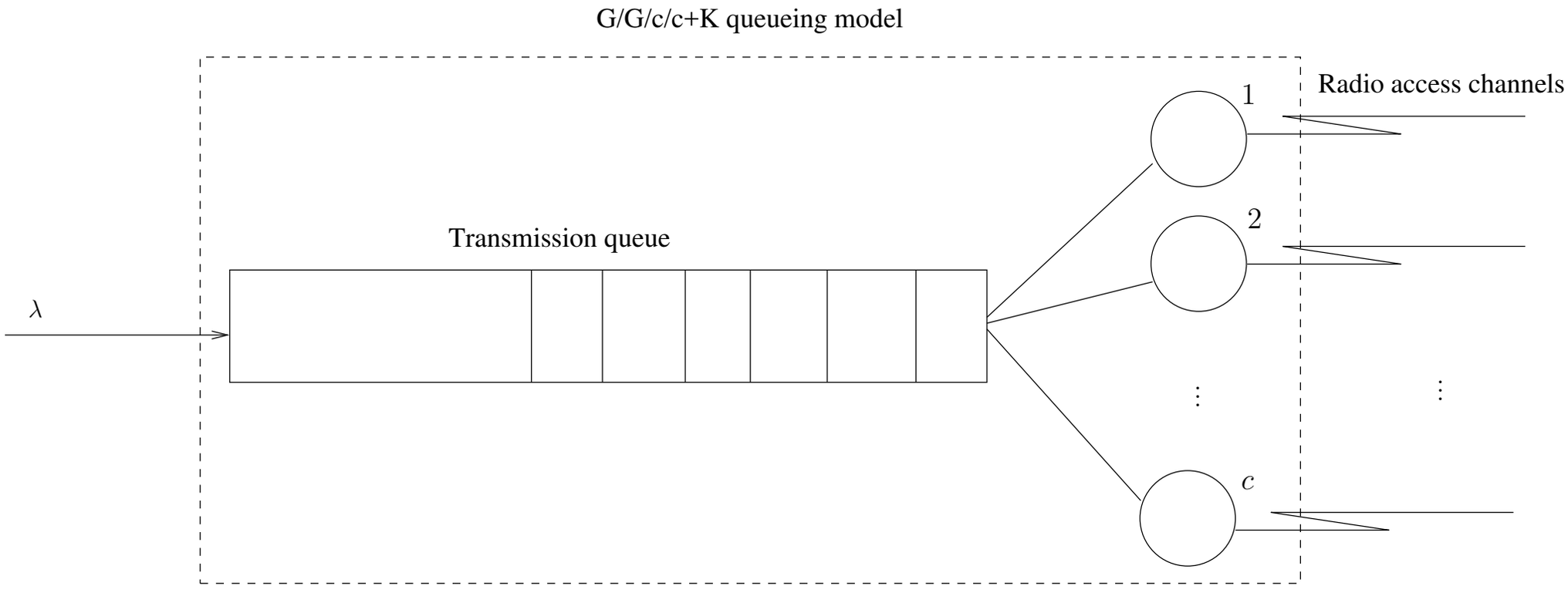


Figure 1: Queueing model for the LTE radio access downlink

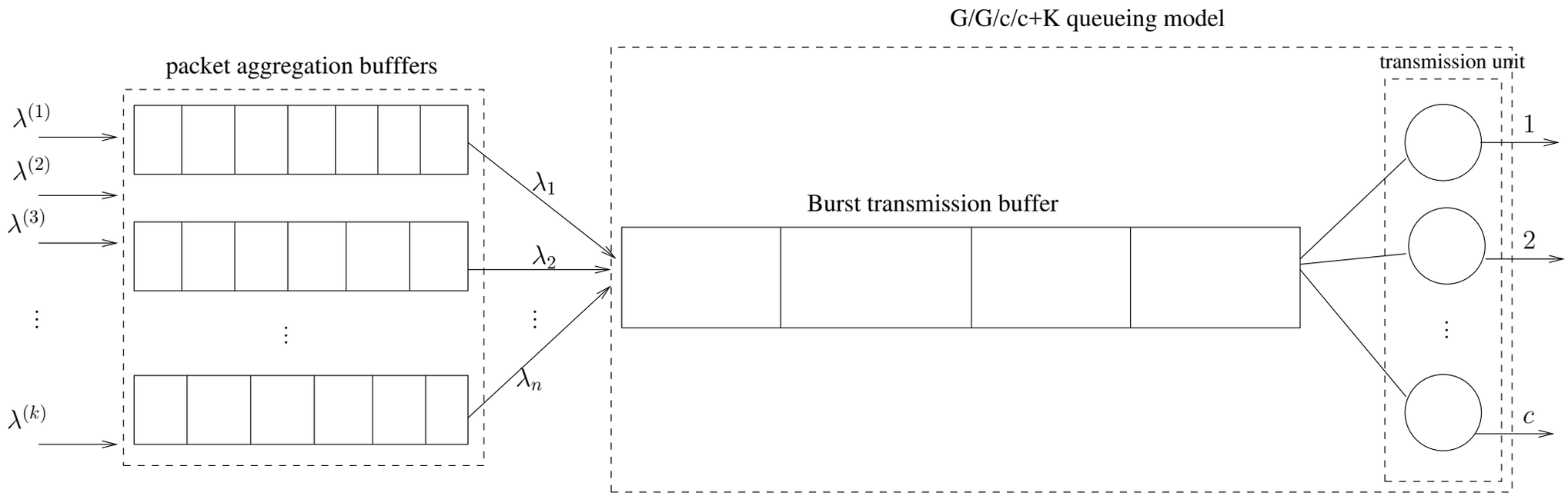


Figure 2: Queueing model of the edge node in optical burst switching networks

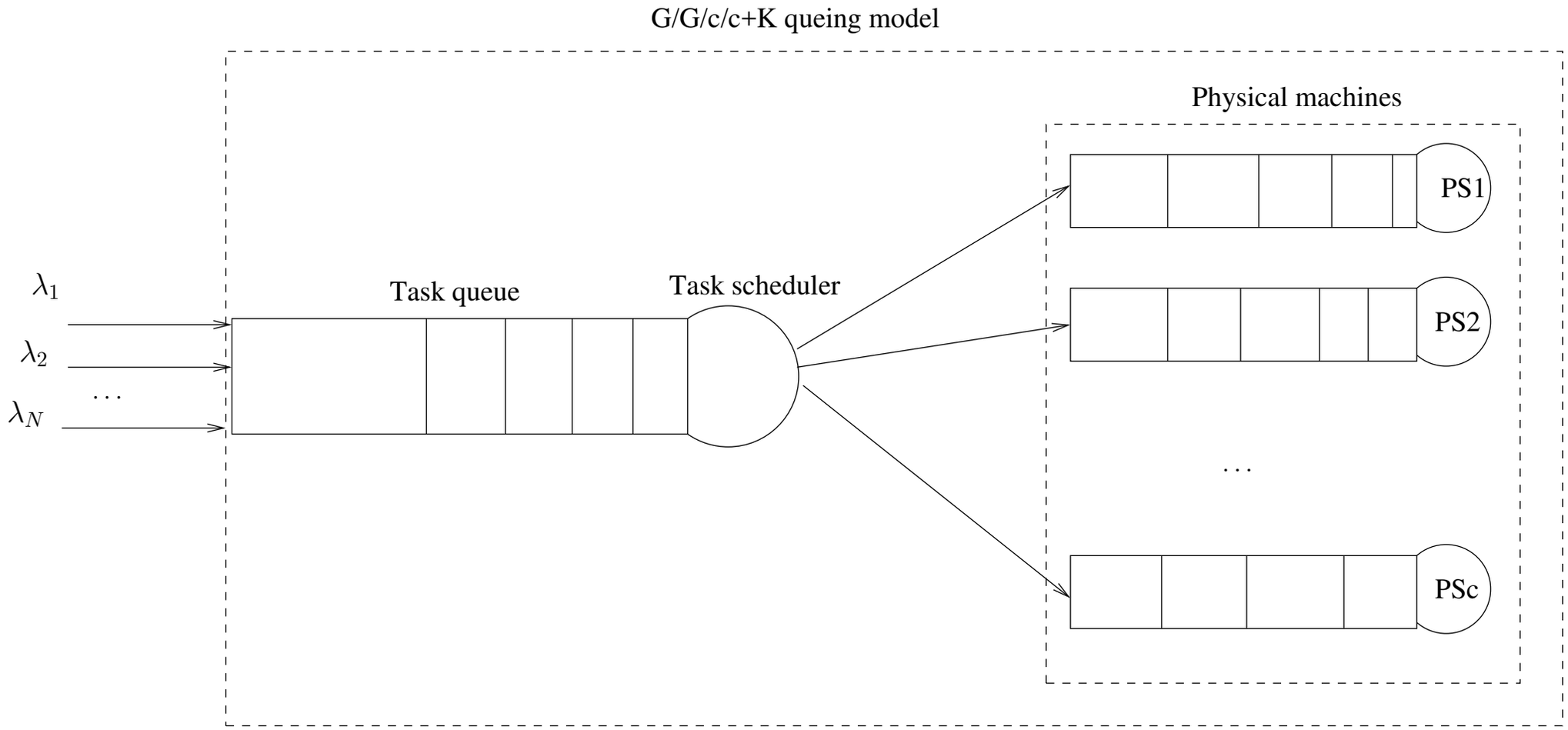


Figure 3: Simplified queuing Model of a Cloud Data Center

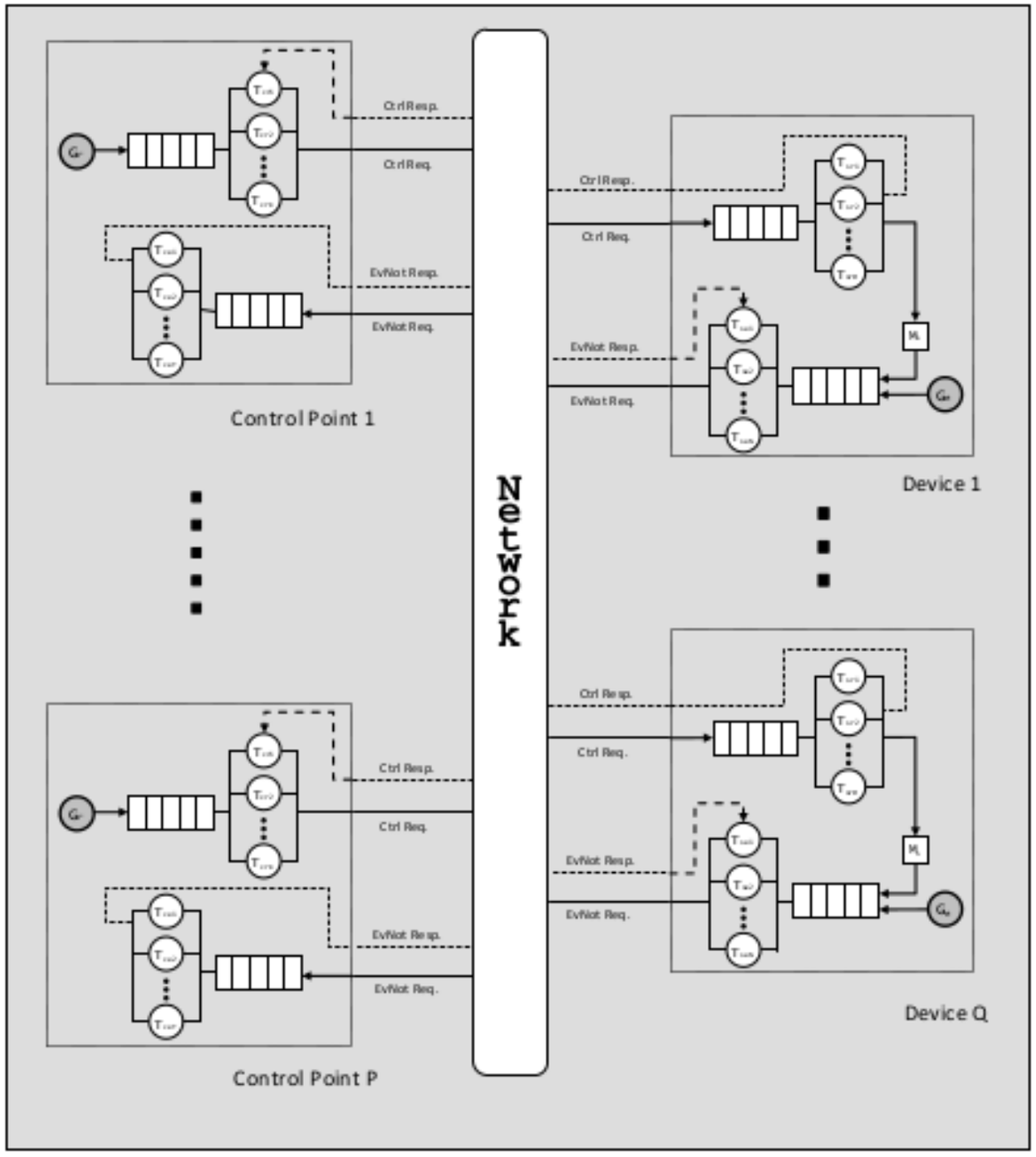


Figure 4: Queuing model of the UPnP/HTTP network

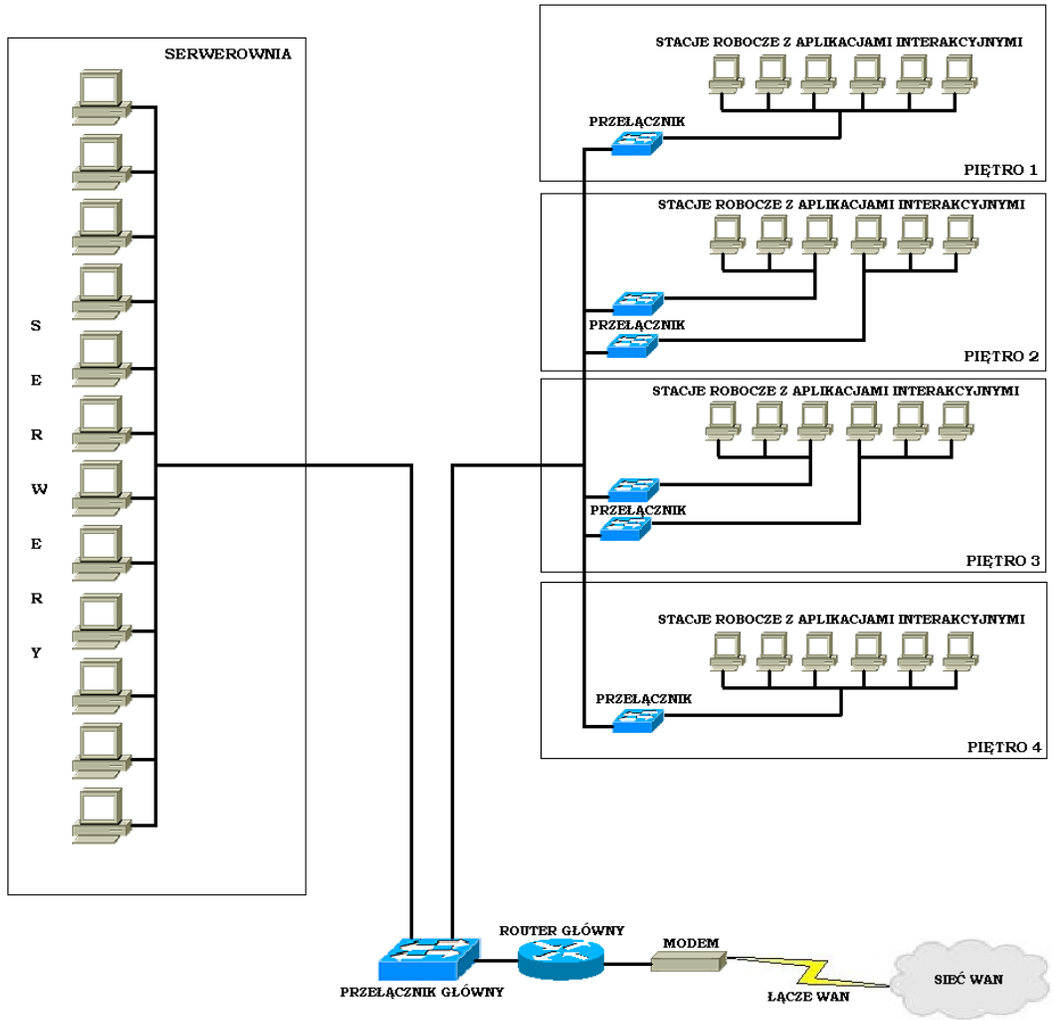


Figure 5: Evaluation of an Polish Insurance (ZUS) Database System

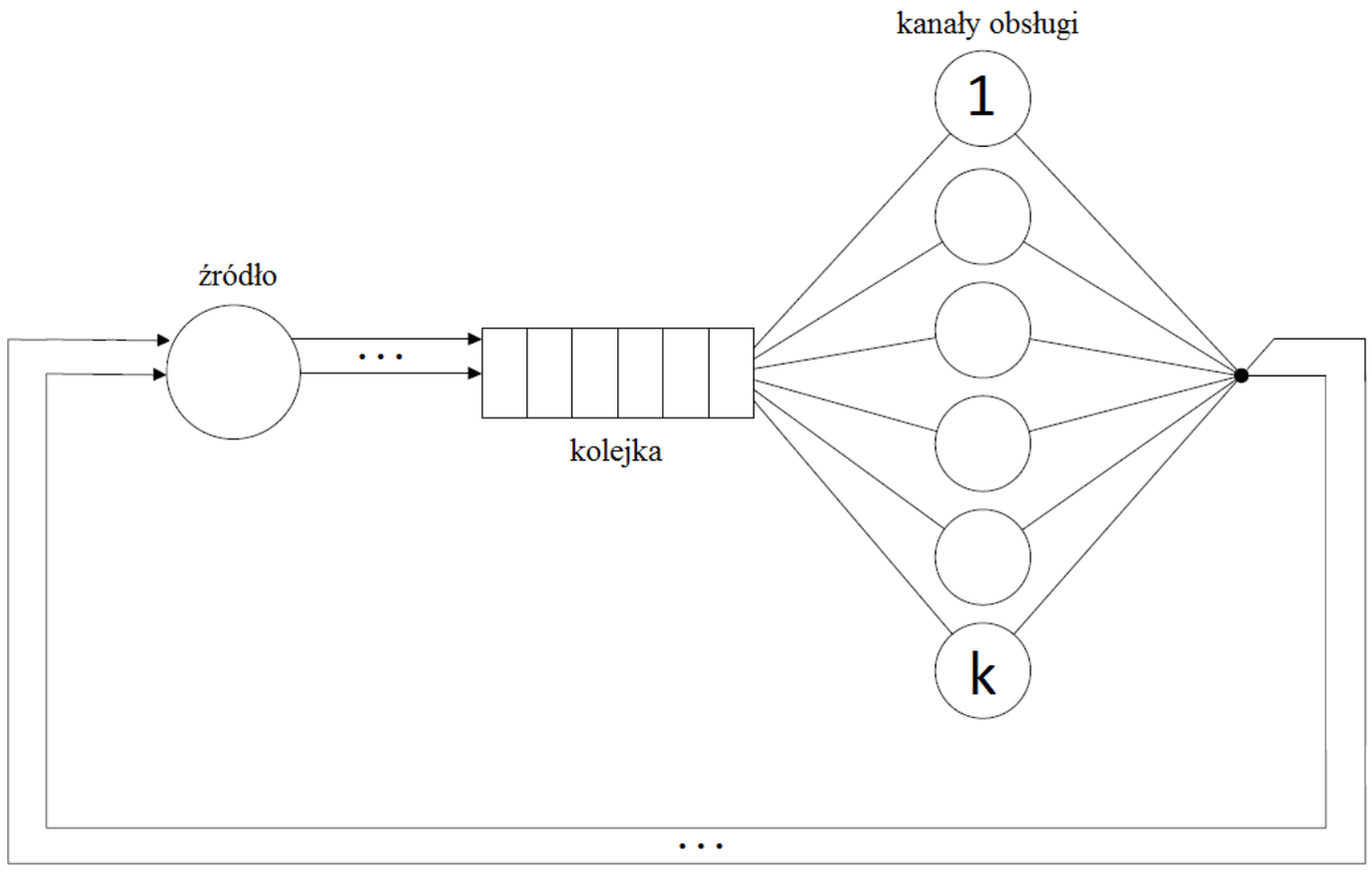


Figure 6: Queuing model of the Database System

Diffusion approximation approach to $G/G/c/c+K$ model

Diffusion approximation is a method introduced to queueing theory by H. Kobayashi and E. Gelenbe in 1970's. Our diffusion G/G/c/c+K model is an extension of Gelenbe's G/G/1/K model presented in [*On approximate computer system models*, E. Gelenbe, Journal of the ACM (JACM), 1975].

We adapt it to multiple channel and finite population case where required diffusion parameters are state-dependent.

We provide also transient state analysis in case of time dependent flows or the change of working channels to see how the queues and blocking probability vary with time and what is the dynamics of the overflow traffic.

The essence of diffusion approximation is the replacement of a stochastic process $N(t)$ – the number of customers in a queueing system by a diffusion process $X(t)$.

The diffusion equation

$$\frac{\partial f(x,t;x_0)}{\partial t} = \frac{\alpha}{2}\frac{\partial^2 f(x,t;x_0)}{\partial x^2} - \beta\frac{\partial f(x,t;x_0)}{\partial x} \ .$$

with appropriate parameters and boundary conditions determines the probability density function $f(x,t;x_0)$ of the process and this function is an approximation of the distribution of the number of customers in the service system

The choice of diffusion parameters

Let $A(x)$, $B(x)$ denote the interarrival and service time distributions. The distributions are general but not specified, the method requires only their two first moments:

means $E[A] = 1/\lambda$, $E[B] = 1/\mu$

and variances $\text{Var}[A] = \sigma_A^2$, $\text{Var}[B] = \sigma_B^2$.

Denote also squared coefficients of variation $C_A^2 = \sigma_A^2\lambda^2$, $C_B^2 = \sigma_B^2\mu^2$.

The changes of $N(t)$ during Δ are normally distributed with mean $(\lambda - \mu)\Delta$, and variance $(\sigma_A^2\lambda^3 + \sigma_B^2\mu^3)\Delta = (C_A^2\lambda + C_B^2\mu)\Delta$.

The changes of $X(t)$ in dt are normally distributed with mean βdt and variance αdt.

Therefore in $G/G/1/N$ model the choice of diffusion parameters is [Gelenbe]

$$\beta = \lambda - \mu,$$

$$\alpha = \sigma_A^2 \lambda^3 + \sigma_B^2 \mu^3 = C_A^2 \lambda + C_B^2 \mu$$

These values assure that the processes $N(t)$ and $X(t)$ have not only normally distributed changes but also their mean and variance increase in the same way with the observation time.

The choice of boundary conditons

In case of $G/G/1/N$ queue, the diffusion process should be limited to the interval $[0, N]$ corresponding to possible number of customers inside the system. To ensure it, two barriers are placed at $x = 0$ and $x = N$.

In Gelenbe's model when the diffusion process comes to $x = 0$, it remains there for a time exponentially distributed with parameter λ and then jumps instantaneously to $x = 1$.

When the diffusion process comes to the barrier at $x = N$ it stays there for a time exponentially distributed with the parameter μ that corresponds to the time for which the queue is saturated and then jumps instantaneously to $x = N - 1$.

The diffusion equation supplemented with jumps and probability balance equations for barriers is

$$\begin{aligned}
\frac{\partial f(x,t;x_0)}{\partial t} &= \frac{\alpha}{2}\frac{\partial^2 f(x,t;x_0)}{\partial x^2} - \beta\frac{\partial f(x,t;x_0)}{\partial x} + \\
&\quad +\lambda_0 p_0(t)\delta(x-1) + \lambda_N p_N(t)\delta(x-N+1)\ , \\
\frac{dp_0(t)}{dt} &= \lim_{x\to 0}\,[\frac{\alpha}{2}\frac{\partial f(x,t;x_0)}{\partial x} - \beta f(x,t;x_0)] - \lambda_0 p_0(t)\ , \\
\frac{dp_N(t)}{dt} &= -\lim_{x\to N}\,[\frac{\alpha}{2}\frac{\partial f(x,t;x_0)}{\partial x} - \beta f(x,t;x_0)] \\
&\quad -\lambda_N p_N(t)\ ,
\end{aligned}$$

where $p_0(t) = P[X(t) = 0]$, $p_N(t) = P[X(t) = N]$.

in the system at time t with initial condition n_0. In case of steady state analysis the equations have the analytical solution and determine $f(x)$, an approximation of $p(n)$ [Gelenbe].

An analytical-numerical method of solution we use is presented in

[*A Method to Solve Diffusion Equation with Instantaneous Return Processes Acting as Boundary Conditions*, T. Czachórski, Bulletin of Polish Academy of Sciences, 1993,]

The value value of $f(n,t;n_0)$ serves as an approximation of $p(n,t;n_0)$, the distribution of the number of customers.

The approach has already a long history, but we adapt it to newly arising problems.

$G/G/c/c+K$ steady state model

The diffusion interval is between two barriers placed at $x = 0$ and $x = c + K$ and is divided into $c - 1$ sub-intervals

$$(0, 1], [1, 2] \ldots [c - 2, c - 1],\ [c - 1, c + K - 1], [c + K - 1, c + K).$$

The last two intervals have the same diffusion parameters but are distinguished because of jumps from $c + K$ to $c + K - 1$. We assume constant parameters inside the sub-intervals having different diffusion parameters

$$\alpha_i = \lambda C_A^2 + i\mu C_B^2, \quad \beta_i = \lambda - i\mu \tag{1}$$

for $i - 1 < x < i, i = 1, 2, \ldots, c - 1$, and

$$\alpha_c = \lambda C_A^2 + c\mu C_B^2, \quad \beta_c = \lambda - c\mu \tag{2}$$

for $c - 1 < x < c + K$.

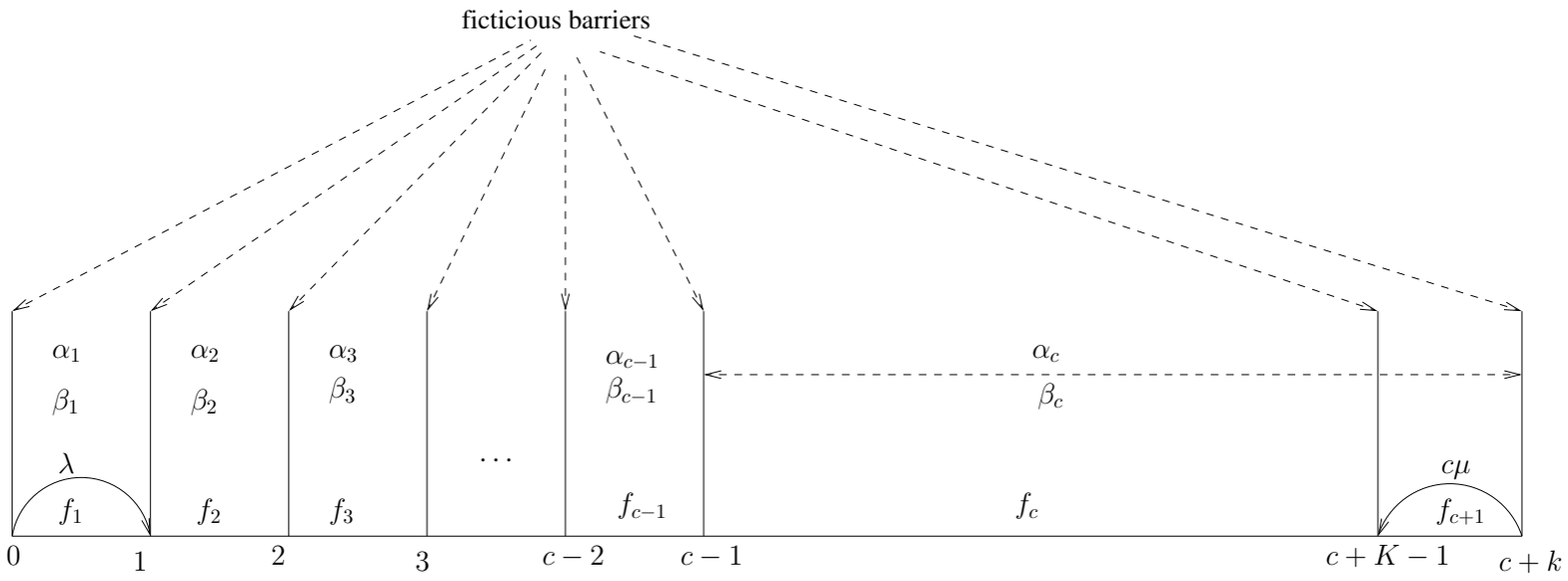


Figure 7: Diffusion intervals and corresponding diffusion parameters

The jumps are performed from $x = 0$ to $x = 1$ with intensity λ and from $x = c + K$ to $x = c + K - 1$ with intensity $c\mu$.

The steady state solution of this equations is

$$f_i(x) = C_{1,i} + C_{2,i} e^{z_i x}, \quad \text{where} \quad z_i = \frac{2\beta_i}{\alpha_i}, \quad i = 1, \dots N.$$

where the constants $C_{1,i}$, $C_{2,i}$ are given by the conditions of continuity of the solution:

$$f_{i-1}(x_i) = f_{i+1}(x_i), \quad i = 1, \dots c + K - 1$$

and the conservation of probability at each sub-interval

$$\frac{\alpha_i}{2} \frac{\partial f_i(x, t; \psi_i)}{\partial x} - \beta f_i(x, t; \psi_i) = 0;$$

in the first and last interval we should include in these balance equations the transport of probability due to jumps.

The additional condition is the normalisation, as the integral od $f_i(x)$ over the interval $[x_{i-1}, x_i]$ is

$$\int_{x_{i-1}}^{x_i} f_i(x)dx = C_{1,i}(x_i - x_{i-1}) + C_{2,i}(1/z_i)[e^{z_i x_i} - e^{z_i x_{i-1}}],$$

then the normalisation equation becomes

$$\begin{aligned} 1 \quad = \quad & p_0 + \sum_{i=1}^{c+K} \{C_{1,i}(x_i - x_{i-1}) + \\ & + C_{2,i}(1/z_i)[e^{z_i x_i} - e^{z_i x_{i-1}}]\} + p_{c+K}. \end{aligned}$$

Using the normalization and continuity condition between subintervals, the steady state solution of the $G/G/c/c+K$ diffusion model with jumps becomes

$$f(x) = \begin{cases} \dfrac{\lambda p_0}{-\beta_1}(1 - e^{z_1 x}), & 0 < x \leq 1 \ , \\ \quad \vdots & \\ \dfrac{\lambda p_0}{-\beta_1}(1 - e^{z_1 x}) e^{z_2(x-1)+\cdots+z_n(x-(n-1))}, & n-1 \leq x \leq n \ , \\ \quad \vdots & \\ \dfrac{\lambda p_0}{-\beta_1}(1 - e^{z_1 x}) e^{z_2(x-1)+\cdots+z_c(x-(c-1))}, & c-1 \leq x \leq c+K-1 \ , \\ \dfrac{c\mu p_{K+c}}{-\beta_m}(e^{z_c(x-(c+K))} - 1), & c+K-1 \leq x < c+K \ , \end{cases} \tag{3}$$

where p_{K+c} is the probability that the diffusion process is at the upper barrier at $x = K + c$ (the queue is saturated), and p_0 is the probability

that the process is at the lower barrier at $x = 0$, i.e. the station is empty,

$$p_{K+c} = \frac{\lambda p_0 \beta_m}{c\mu\beta_1} \left[\frac{1 - e^{z_1(c+K-1)}}{e^{-z_c} - 1} \right] e^{z_2(c+K-2)+\cdots+z_c K}$$

where $\rho = \lambda/c\mu$ and p_0 is determined from the normalization condition.

Waiting time and response time distributions

If the number of customers $n < c$, there is no waiting time, the response time is just service time. If $n \geq c$ the waiting time for the end of service of $n - c + 1$ customers. Because c service channels are active, the time to end the nearest service may be computed as

$$F_{Bc}(x) = 1 - (1 - F_B(x))^c$$

and

$$f_{Bc}(x) = \frac{dF_{Bc}(x)}{dx} = c(1 - F_B(x))^{c-1} f_B(x) \tag{4}$$

where

- $f_{Bc}(x)$ - probability density function (pdf) of the time till the end of the nearest service,
- $F_{Bc}(x)$ - probability distribution function (PDF) of this time
- $F_B(x)$ - probability distribution function of the service time;

The waiting time has the pdf $f_W(t)$

$$\begin{aligned} f_W(x) &= [p(0) + \cdots p(c-1)]\,\delta(x) + p(c) f_{Bc}(x) + \\ &\quad + p(c+1) f_{Bc}^{*2}(x) + \cdots + \\ &\quad p(c+K-1) f_{Bc}^{*(c+K-1)}(x) \end{aligned}$$

where * denotes the convolution and $*i$ is i-fold convolution. The response time is the sum of waiting time and service time, hence its pdf $f_R(x)$ is

$$f_R(x) = f_W(x) * f_B(x)$$

The probability density function (PDF) of of the first passage time is

$$\gamma(t, x_0) = \frac{x_0}{\sqrt{2\pi\alpha t^3}} e^{-\left[\frac{2x_0\beta}{\alpha} + \frac{(x_0-\beta)^2}{2\alpha t}\right]} \tag{5}$$

where β and α are the parameters of the homogenous diffusion movement. Then the density $f_W(t)$ of the waiting time may be expressed as the density of first passage time from any point $\xi \in [c, c+K]$ taken with density $f(\xi)$ of the queue length distribution, provided that the number of customers exceed c and there is waiting time to $x = c$

$$f_W(t) = \frac{\int_{c-1}^{c+K} \gamma(t, \xi) f(\xi) d\xi}{\int_{c-1}^{c+K} f(\xi) d\xi} \tag{6}$$

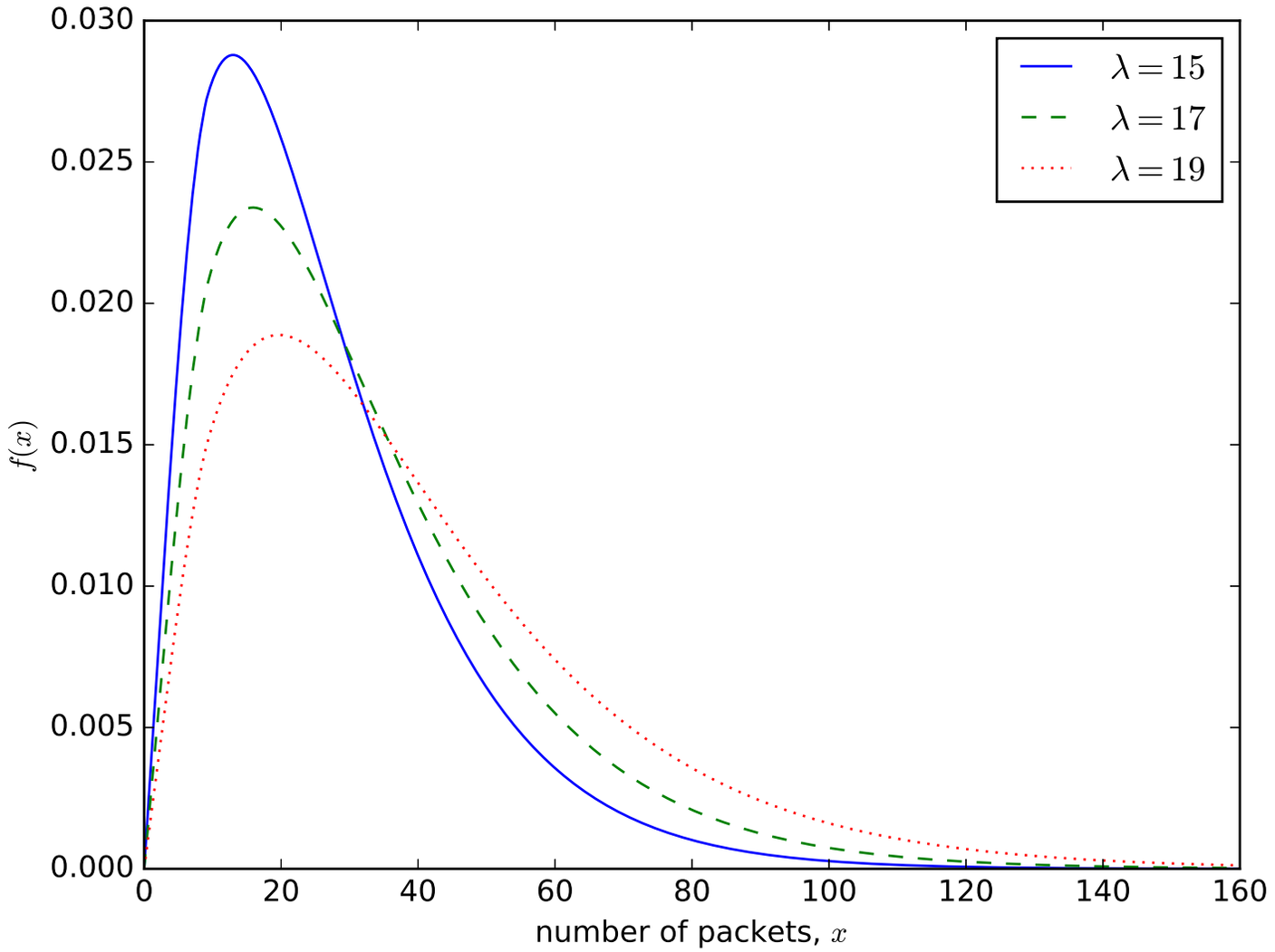


Figure 8: The influence of input traffic rate λ on the distribution of the number of packets for $K = 150$, $c = 10$ and $\mu = 10$

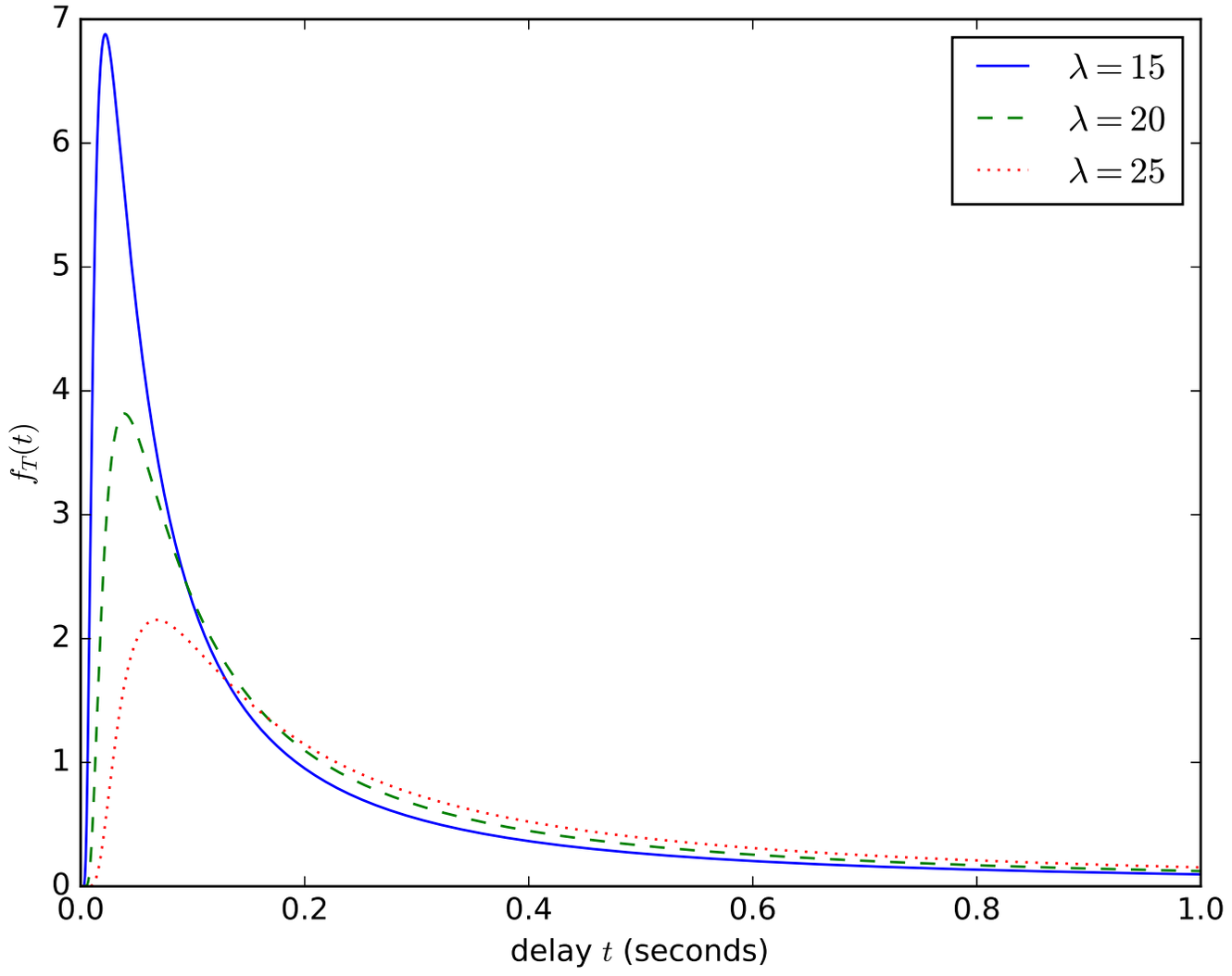


Figure 9: The influence of input traffic rate λ on the distribution of the delay (response time) for $K = 150$, $c = 10$ and $\mu = 10$

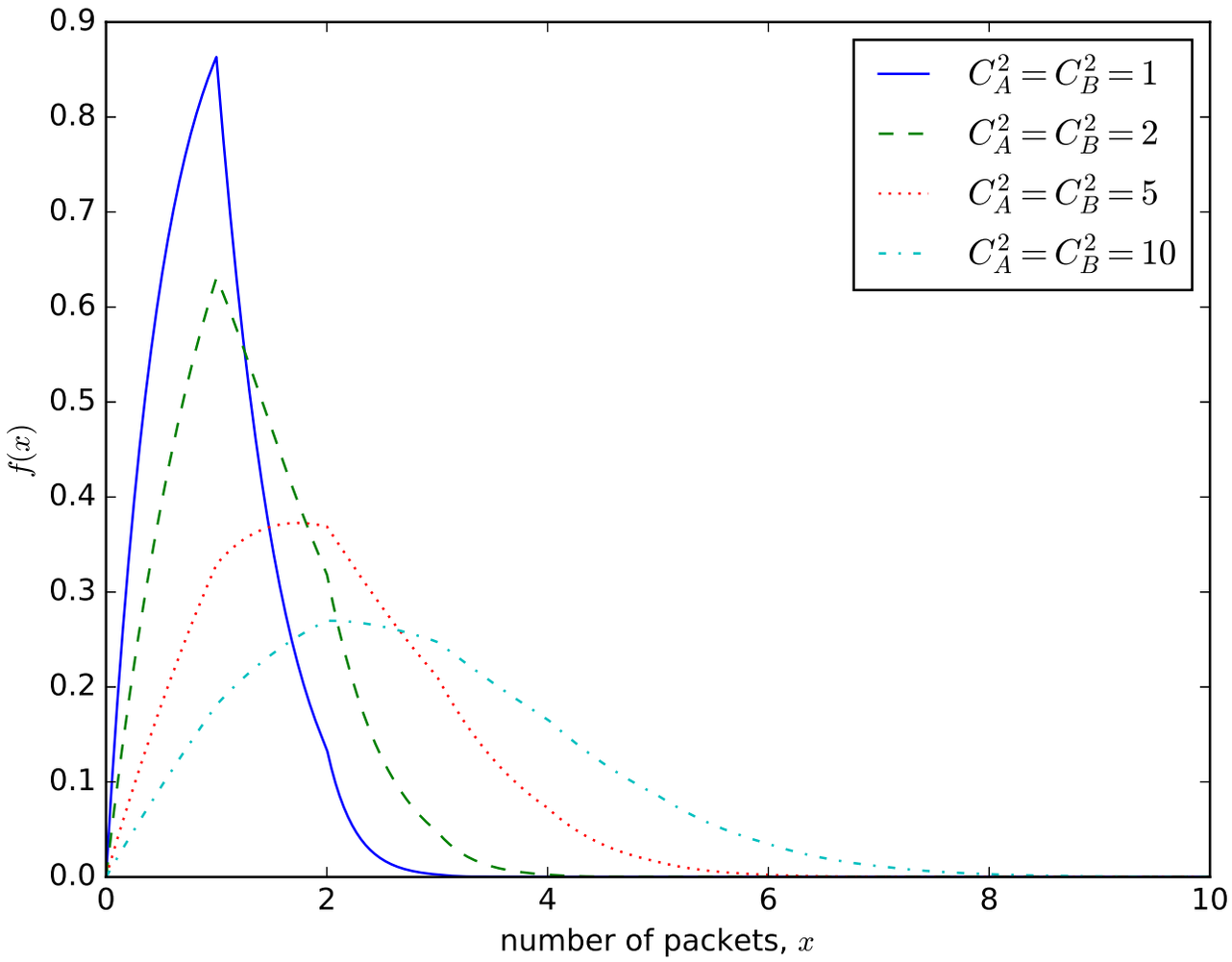


Figure 10: The influence of the squared coefficient of variation of the arrival and service times C_A^2, C_B^2, on the distribution of the number of customers for $K = 50$, $c = 10$ and $\mu = 30$, $\lambda = 20$

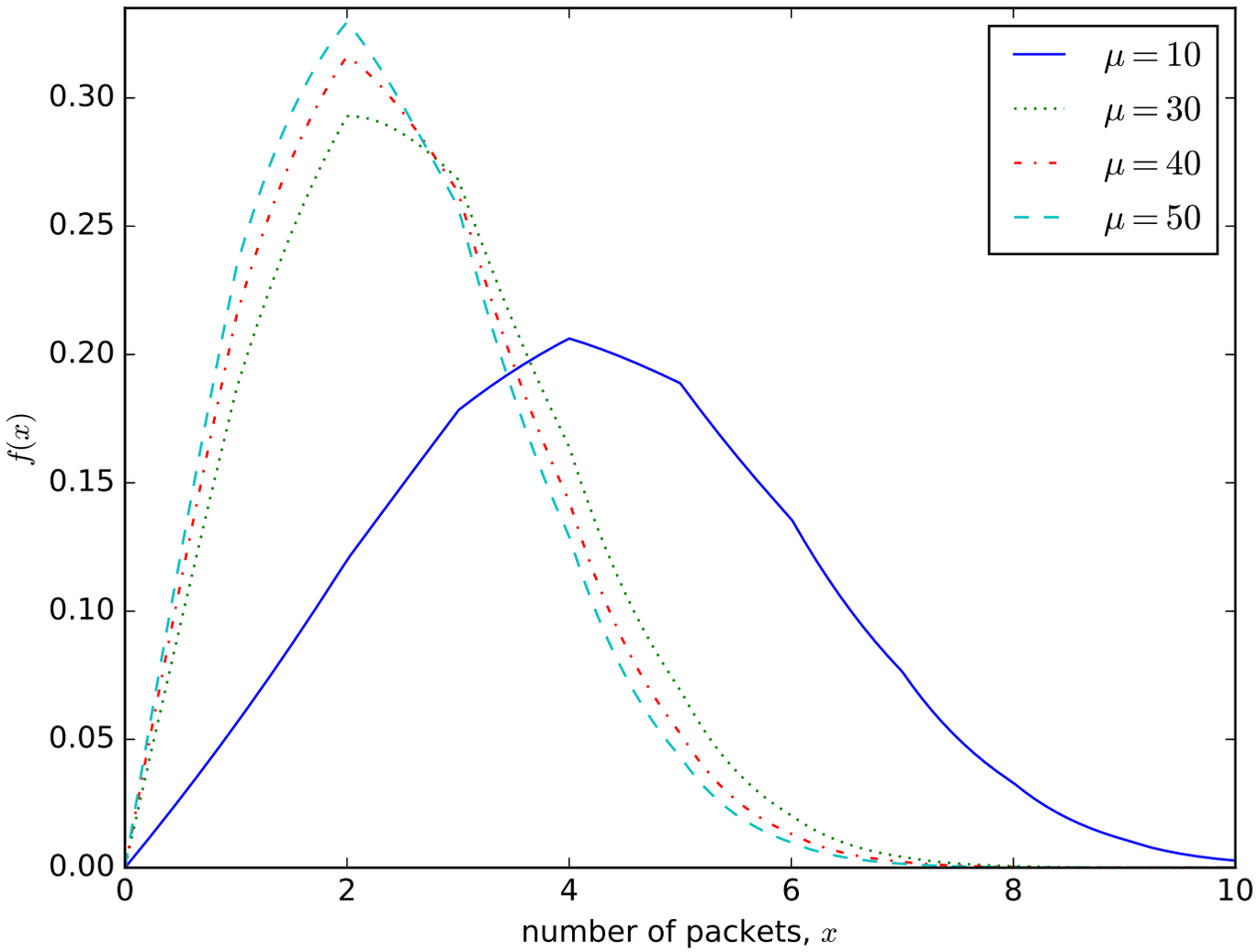


Figure 11: The influence of the service rates μ on the distribution of the number of packets for $K = 50$, $\lambda = 20$ and $c = 10$, $C_A^2 = C_B^2 = 5$

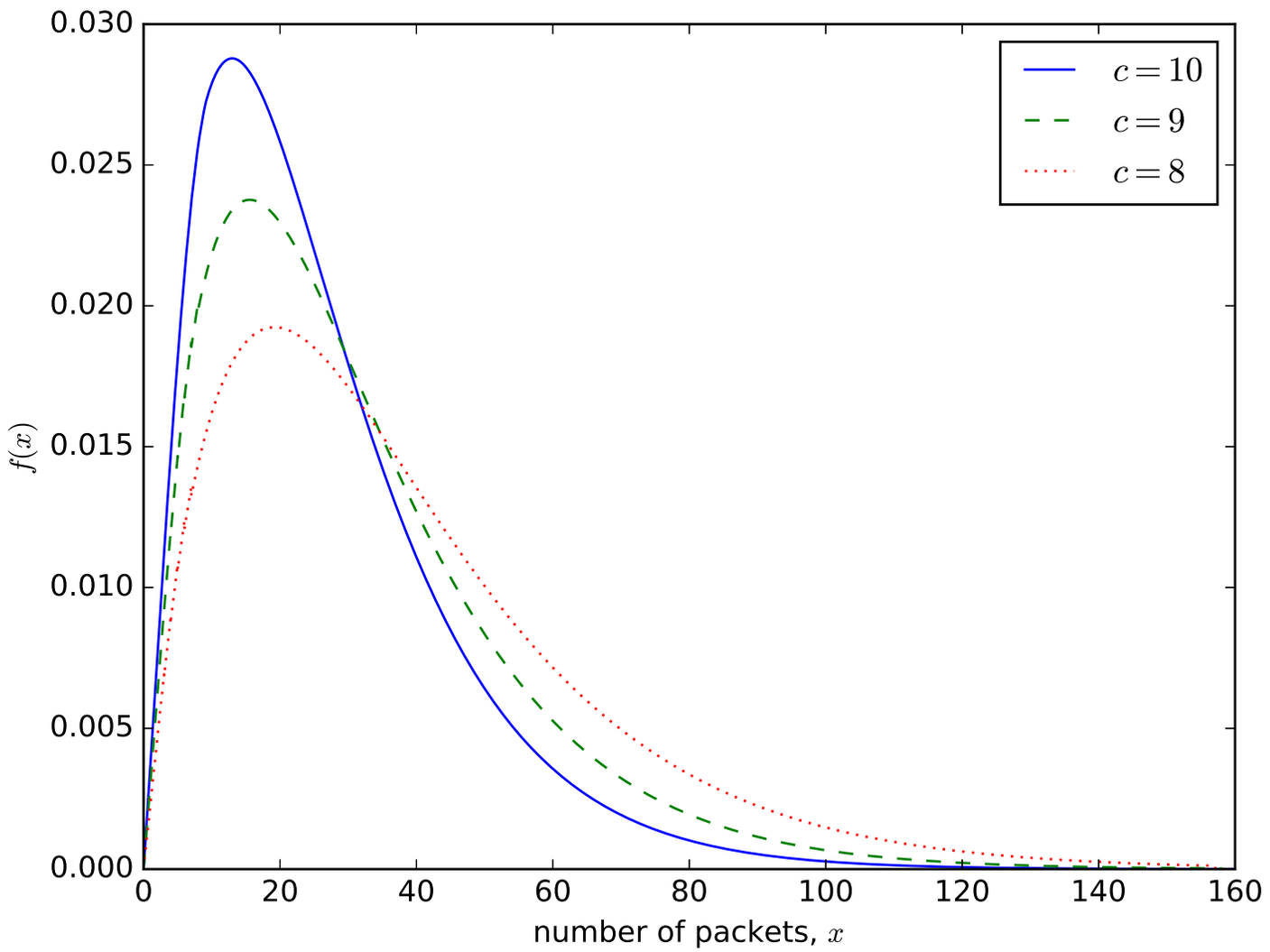


Figure 12: The influence of the number of channels c on the distribution of the number of customers for $K = 150$, $\lambda = 15$ and $\mu = 10$

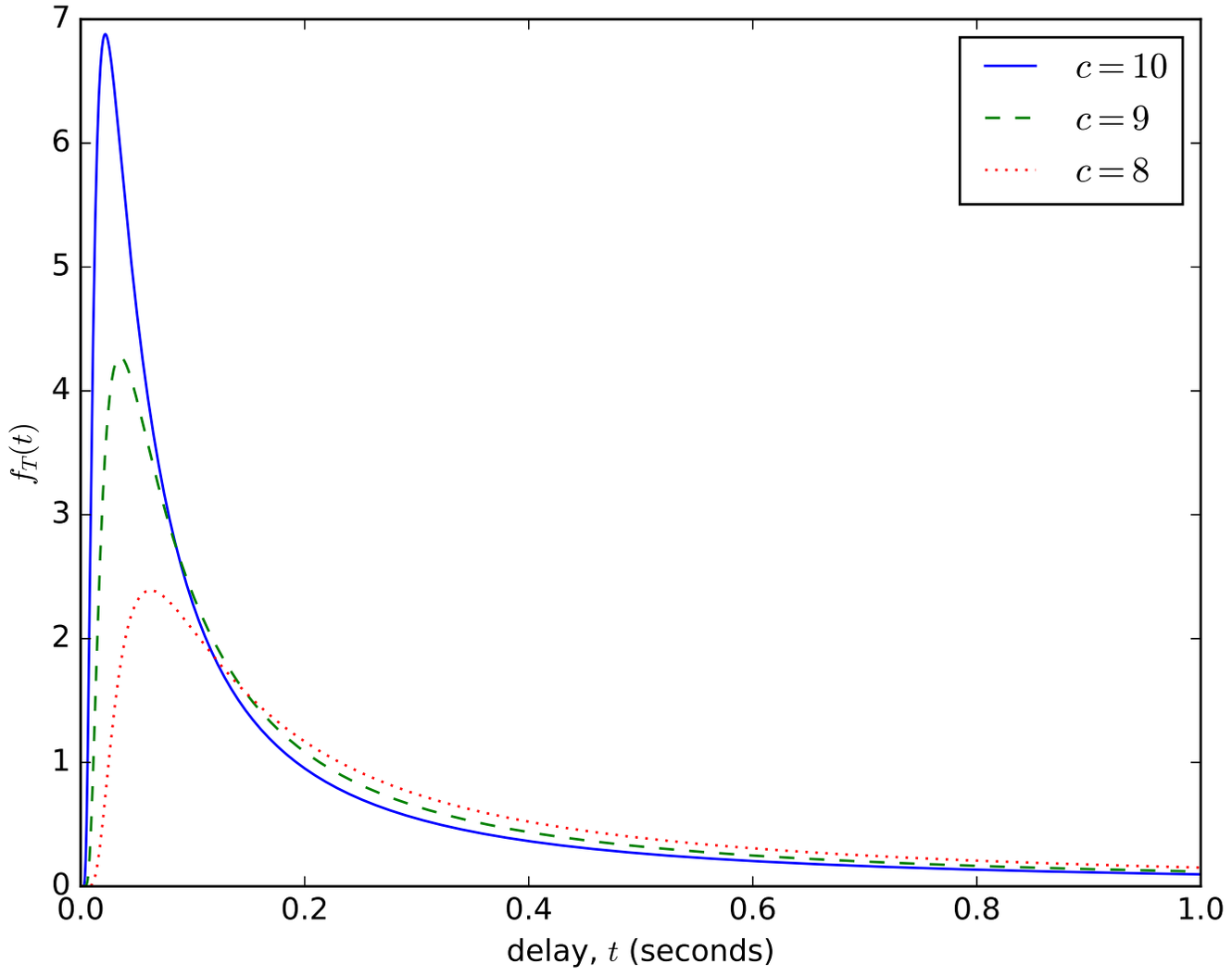


Figure 13: The influence of the number of service channels c on the distribution of the delay (response time) for $K = 150$, $\lambda = 15$ and $\mu = 10$

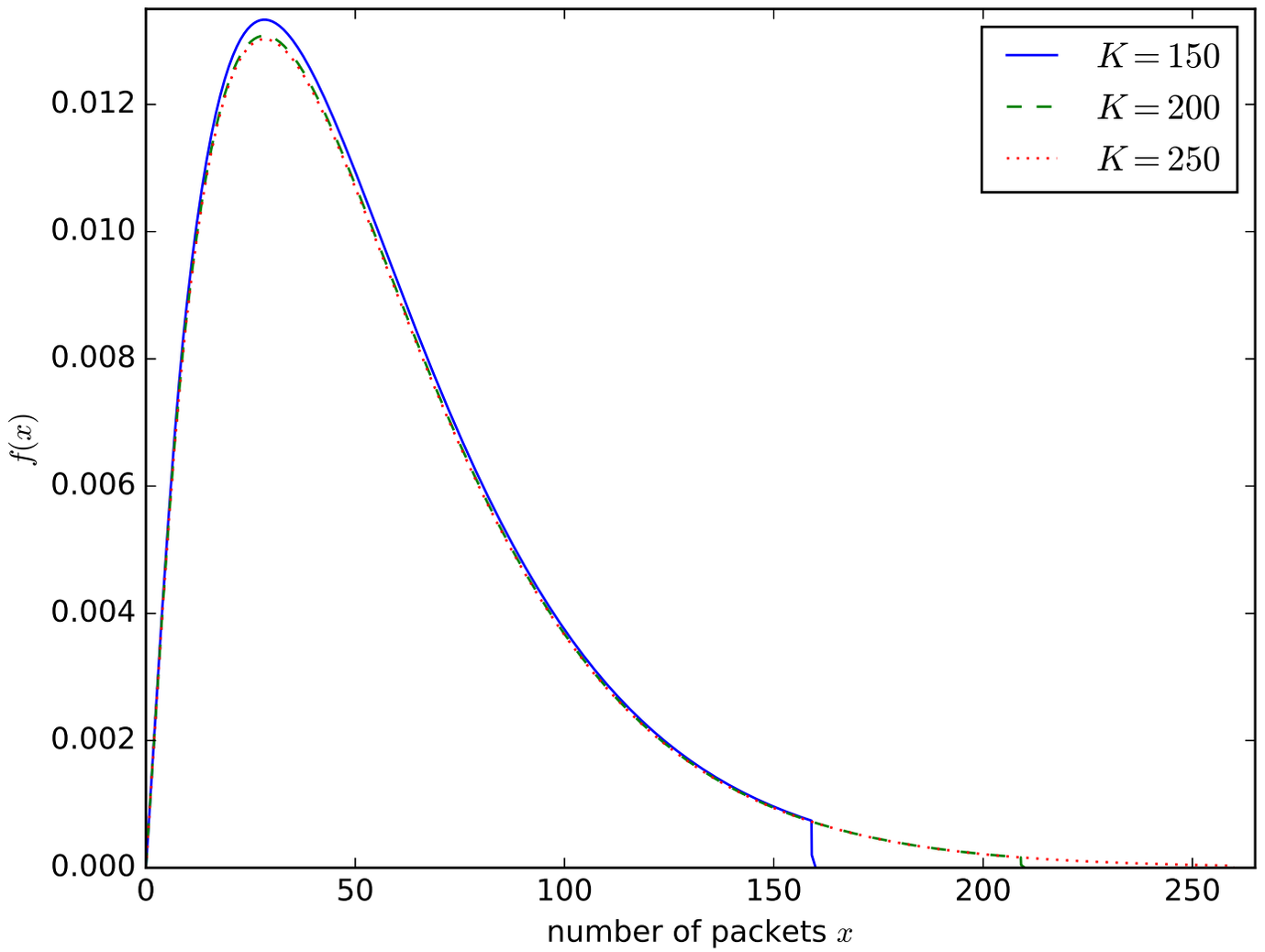


Figure 14: The influence of the buffer size K on the distribution of the number of packets for $\lambda = 15$ and $c = 10$, $\mu = 20$

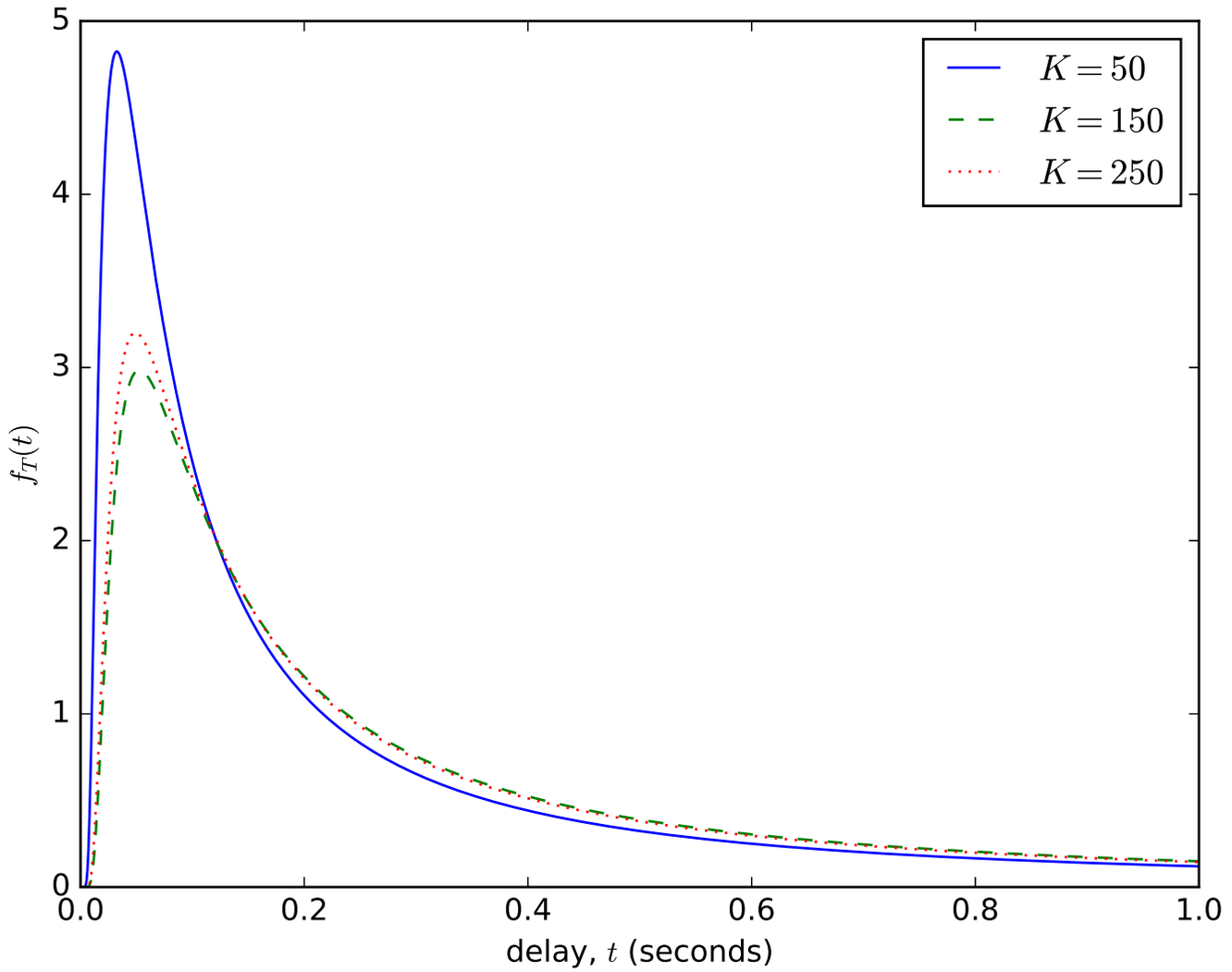


Figure 15: The influence of the buffer size K on the distribution of the delay (response time) for $\lambda = 15$ and $c = 10$, $\mu = 20$

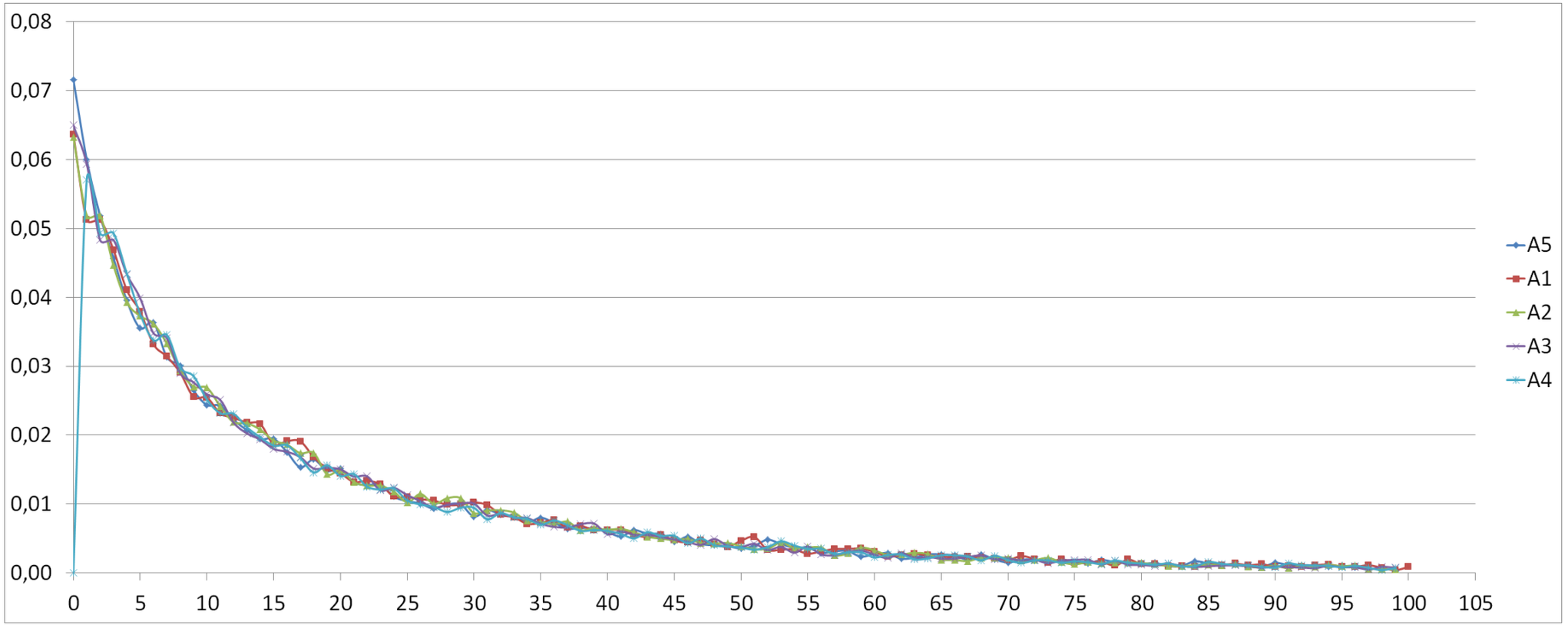


Figure 16: Applications $A_1 \cdots A_5$, histogram of execution times

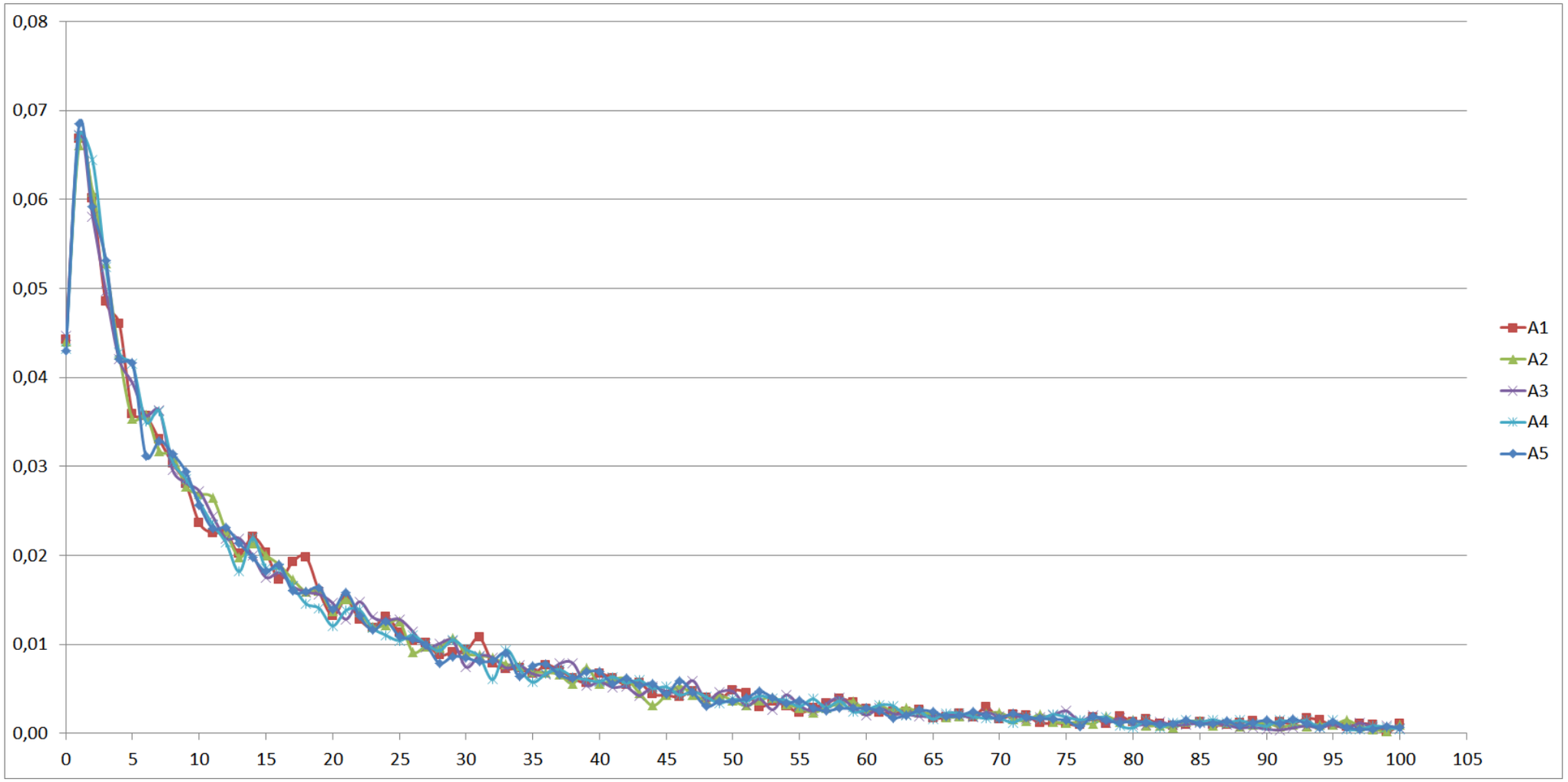


Figure 17: Applications $A_1 \cdots A_5$, histogram of the times in source

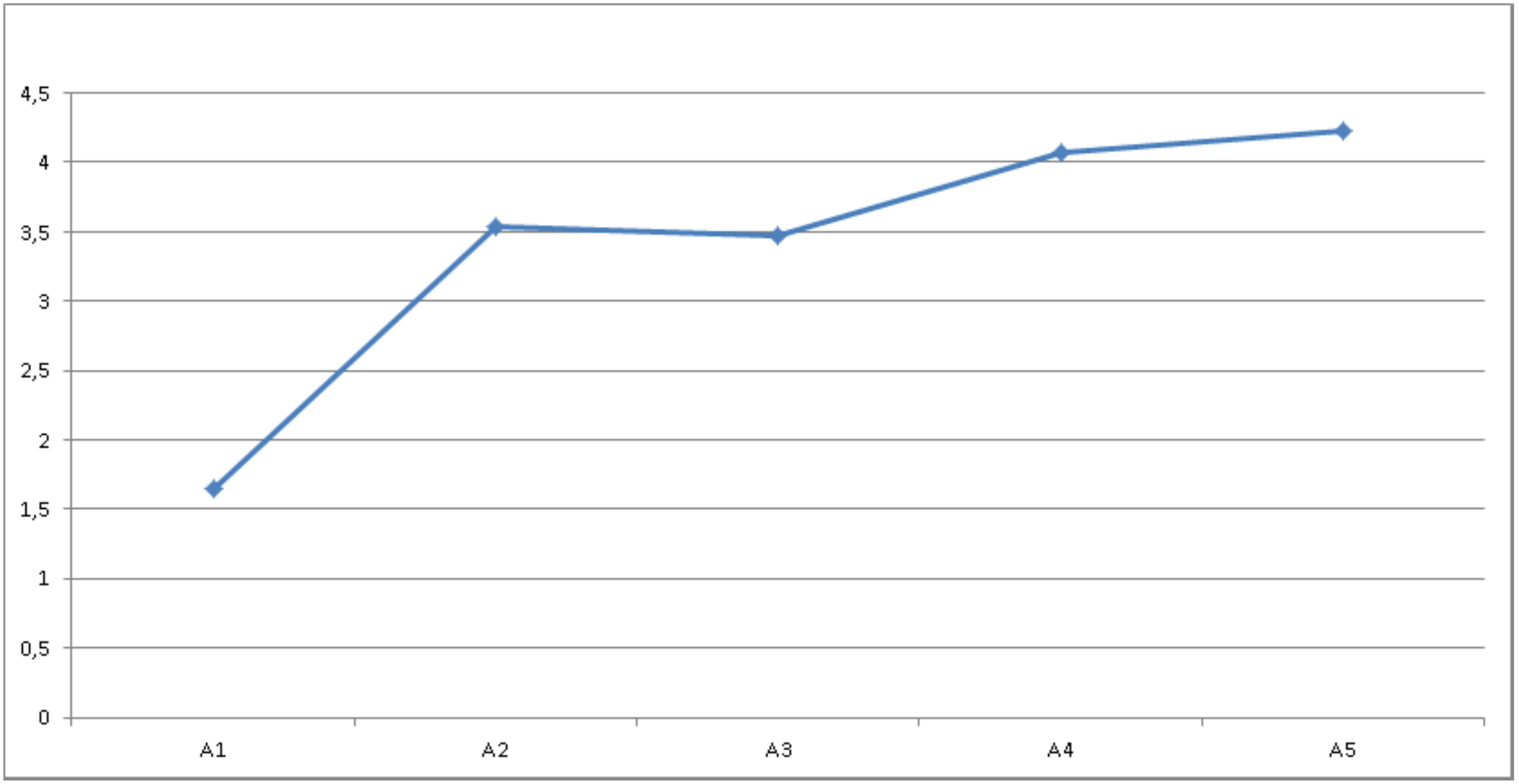


Figure 18: Applications $A_1 \cdots A_5$, sqaured coefficient of variation of the times in source

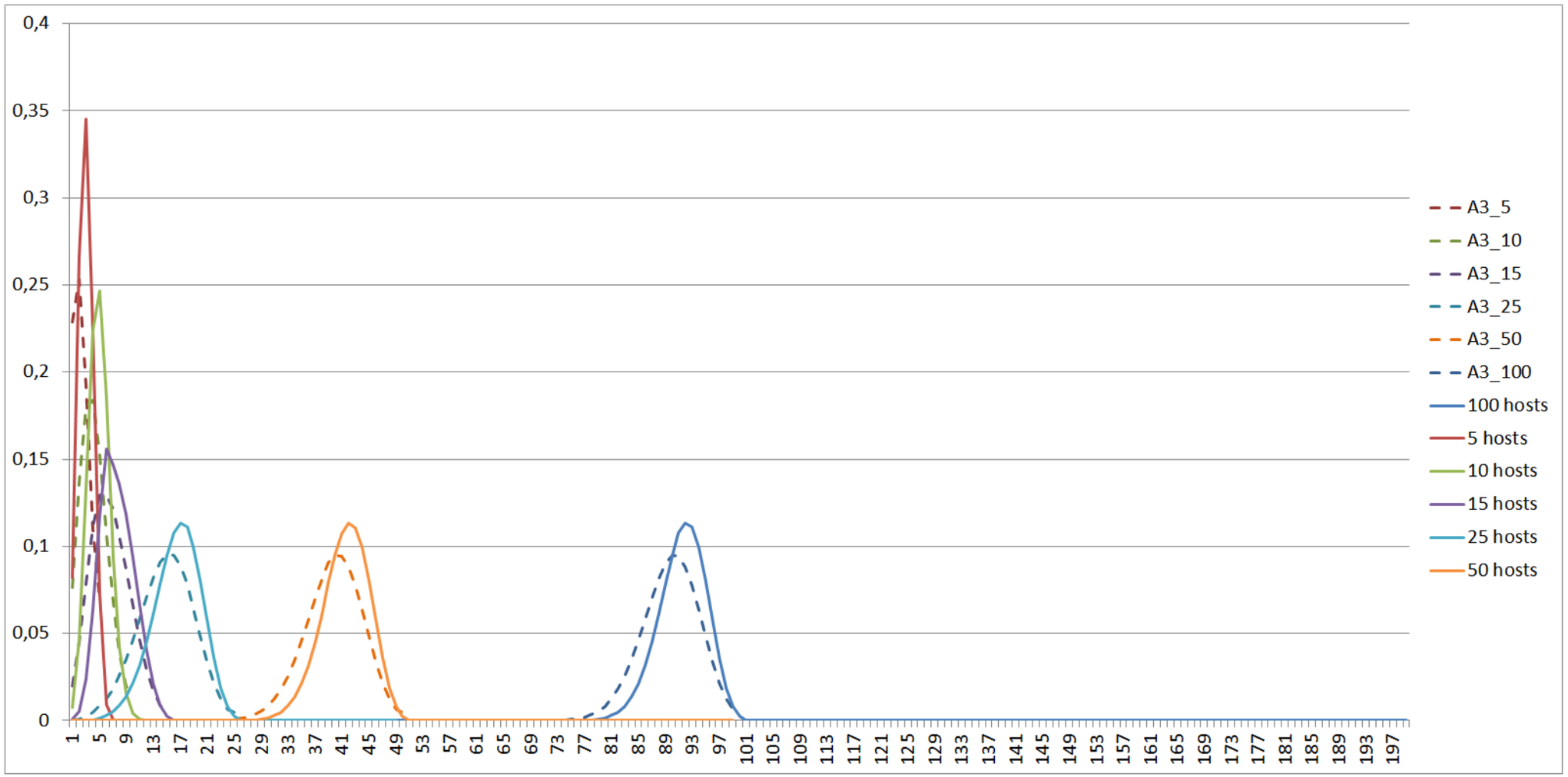


Figure 19: G/G/6//H, set of measured service times A_4, $f(x)$ by diffusion and $p(n)$ by simulation as a function of the number of active clients H.

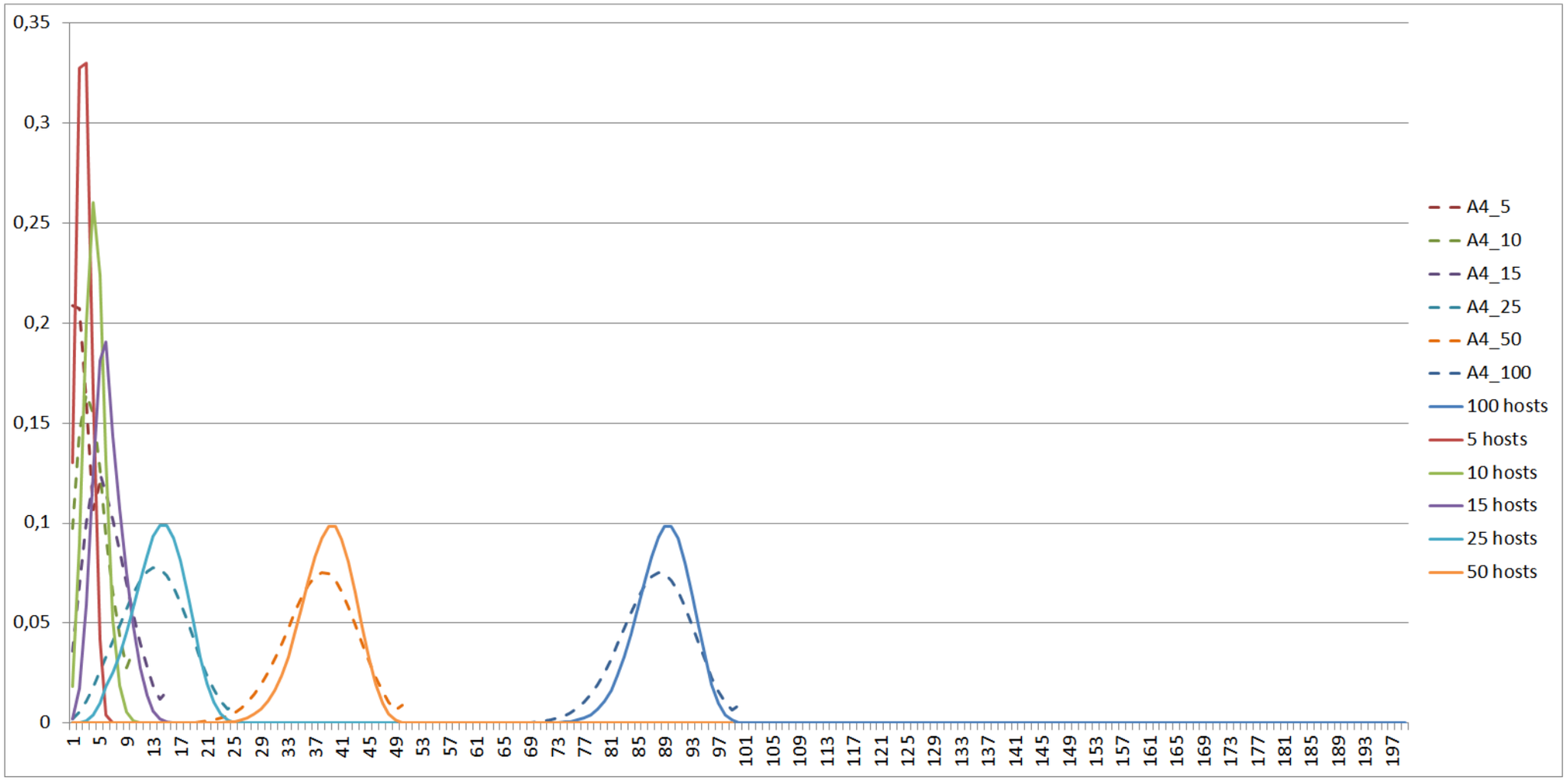


Figure 20: G/G/6//H, set of measured service times A_5, $f(x)$ by diffusion and $p(n)$ by simulation as a function of the number of active clients H.

Transient solution

There is no probability transfer between intervals in steady-state solution but we should take it into account in transient. Inside each of intervals, the diffusion equation is solved assuming that the barriers at its left and right side act as absorbing ones. The density function $\phi(x, t; x_0)$ of a diffusion process limited by two absorbing barriers at $x = 0$ and $x = N$ and with the initial condition $x = x_0$ at $t = 0$ is, see e.g. [Cox-Miller-1965],

$$\phi(x, t; x_0) = \frac{1}{\sqrt{2\Pi\alpha t}} \sum_{n=-\infty}^{\infty} (a_n - b_n)$$

where

$$
\begin{aligned}
a_n &= \exp\left[\frac{\beta x'_n}{\alpha} - \frac{(x - x_0 - x'_n - \beta t)^2}{2\alpha t}\right], \\
b_n &= \exp\left[\frac{\beta x''_n}{\alpha} - \frac{(x - x_0 - x''_n - \beta t)^2}{2\alpha t}\right]
\end{aligned}
$$

and $x'_n = 2nN,\ x''_n = -2x_0 - x'_n$.

To balance probability flows between neighboring intervals having different diffusion parameters, we put fictitious barriers between these intervals and suppose that the diffusion process which is entering a barrier at $x = i$, from its left side (the process is increasing) is absorbed and immediately reappears at $x = i + \varepsilon$. Similarly, a process which is decreasing and enters the barrier from its right side reappears at its other side at $x = n - \varepsilon$. The value of ε should be small, for example of the order of 2^{-10}, but we checked that it has no significant impact on the solution.

The density function $f_i(x,t;\psi_i)$, for an interval i (x_{i-1},x_i) is expressed as

$$\begin{aligned} f_i(x,t;\psi_i) &= \phi_i(x,t;\psi_i) + \\ &\int_0^t g_{x_{i-1}+\varepsilon}(\tau)\phi_i(x,t-\tau;x_{i-1}+\varepsilon)d\tau + \\ &\int_0^t g_{x_i-\varepsilon}(\tau)\phi_n(x,t-\tau;x_i-\varepsilon)d\tau. \end{aligned} \tag{7}$$

The system of equations defining all $f_i(x,t;\psi_i)$ with the use of flows appearing at the vicinity of the barriers, defined by flows entering the barriers from neighboring intervals and expressed with the use of $f_{i-1}(x,t;\psi_{i-1})$ and $f_{i+1}(x,t;\psi_{i+1})$:

$$
\begin{aligned}
g_{x_{i-1}+\varepsilon}(t) &= \lim_{x \to x_{i-1}} [\frac{\alpha_{i-1}}{2} \frac{\partial f_{i-1}(x,t;\psi_{i-1})}{\partial x} - \\
&\qquad \beta_{i-1} f_{i-1}(x,t;\psi_{i-1})] \\
g_{x_i-\varepsilon}(t) &= -\lim_{x \to x_i} [\frac{\alpha_{i+1}}{2} \frac{\partial f_{i+1}(x,t;\psi_{i+1})}{\partial x} - \\
&\qquad \beta_{i+1} f_{i-1}(x,t;\psi_{i+1})]
\end{aligned}
$$

is transformed with the use of Laplace transform and solved numerically to obtain the values of $\bar{f}_i(x,s;\psi_i)$. Then we use the Stehfest inversion algorithm to compute $f_i(x,t;\psi_i)$; for a specified t. This solution gives transient behaviour od the considered system but the parameters of the model do not vary with time.

However, we are interested in **time-dependent input flow**, hence the model is applied to small time-intervals, typically of the length of one mean service time, where the parameters are considered constant and the solution at the end of each time interval gives the initial conditions for the next one.

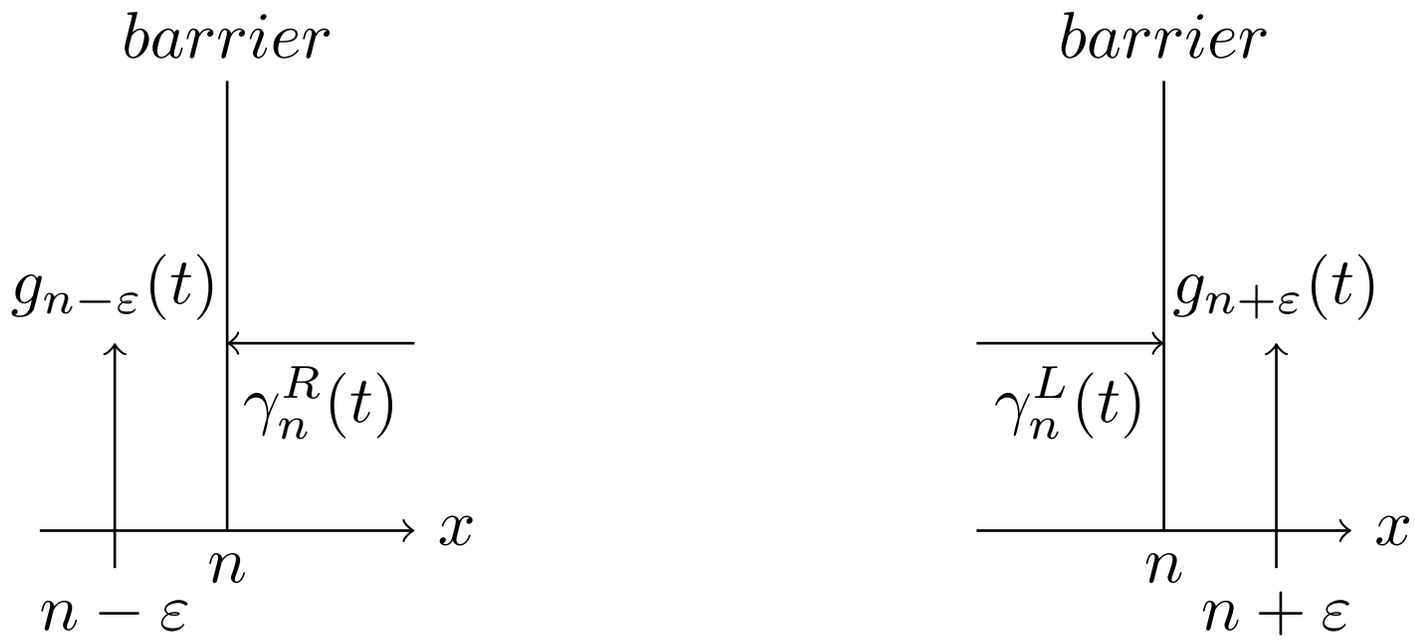


Figure 21: Flow balance for the barrier at $x = n$

The density functions for the intervals are as follows:

$$
\begin{aligned}
f_1(x,t;\psi_1) &= \phi_1(x,t;\psi_1) + \int_0^t g_1(\tau)\phi_1(x,t-\tau;1)d\tau + \\
&\quad + \int_0^t g_{1-\varepsilon}(\tau)\phi_1(x,t-\tau;1-\varepsilon)d\tau\,, \\
f_n(x,t;\psi_n) &= \phi_n(x,t;\psi_n) + \int_0^t g_{n-1+\varepsilon}(\tau)\phi_n(x,t-\tau;n-1+\varepsilon)d\tau + \\
&\quad + \int_0^t g_{n-\varepsilon}(\tau)\phi_n(x,t-\tau;n-\varepsilon)d\tau, \quad n=2,\dots c+K-1, \\
f_{c+K}(x,t;\psi_{c+K}) &= \phi_{c+K}(x,t;\psi_{c+K}) + \\
&\quad \int_0^t g_{c+K-1+\varepsilon}(\tau)\phi_{c+K}(x,t-\tau;c+K-1+\varepsilon)d\tau + \\
&\quad + \int_0^t g_{c+K-1}(\tau)\phi_{c+K}(x,t-\tau;c+K-1)d\tau \qquad (8)
\end{aligned}
$$

This approach is mastered numerically and the errors of the approximation were studied for various models.

This approach may be applied to a **network** of G/G/c/c+K stations, following the principles of decomposition of a G/G/1 or G/G/1/N network presented in [Gelenbe].

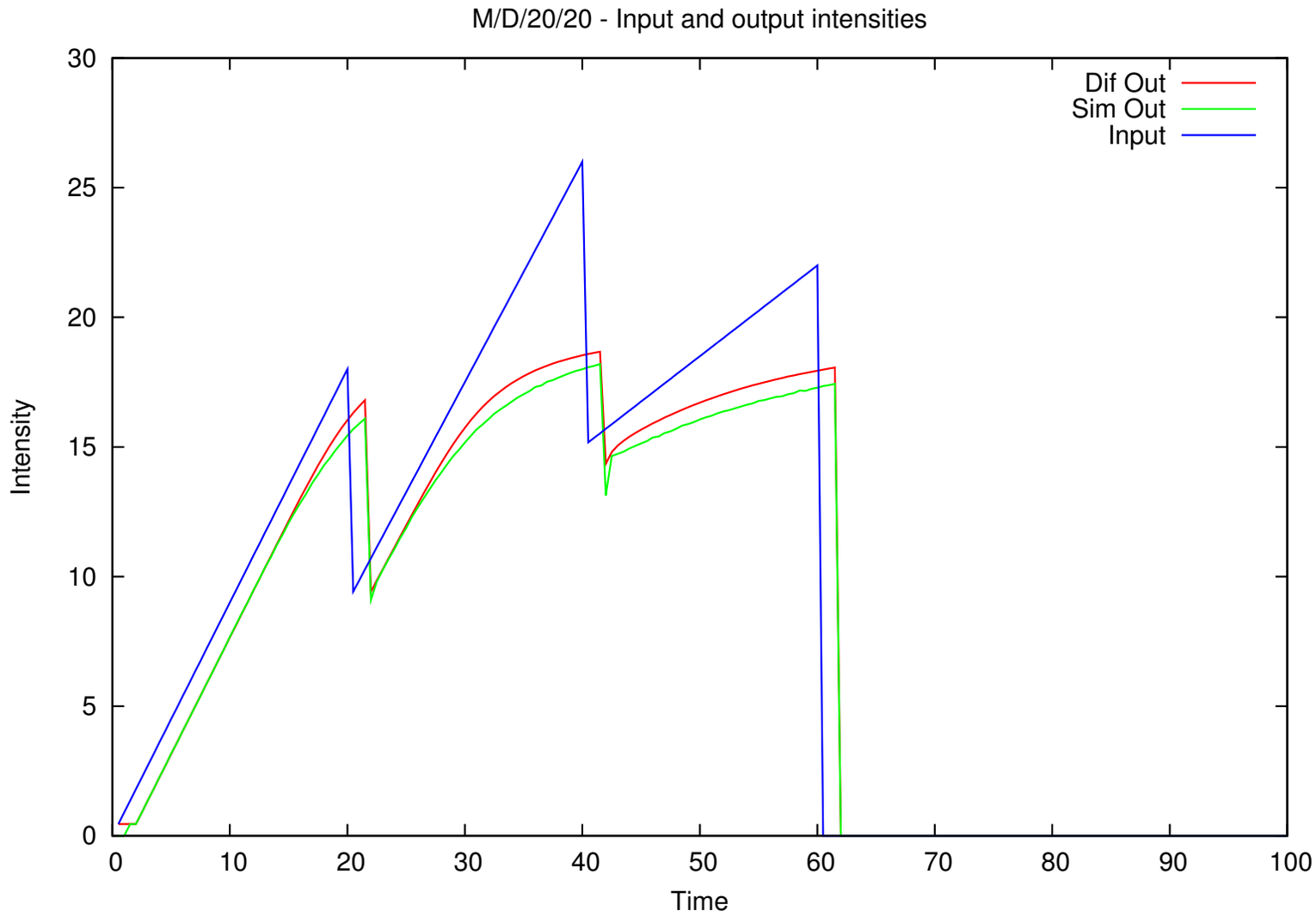


Figure 22: Input stream intensity λ and output streams intensities

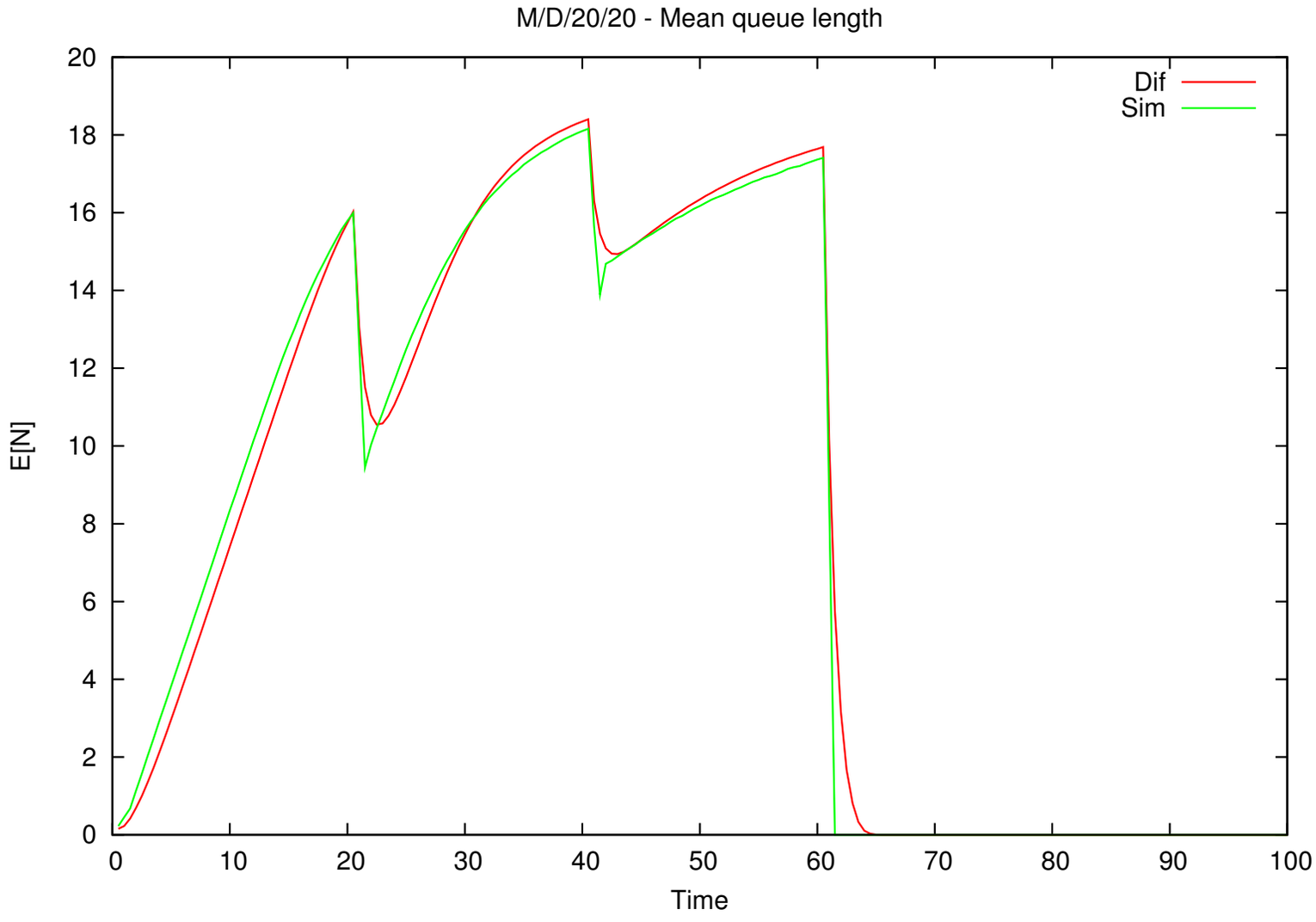


Figure 23: Mean number of customers as a function of time

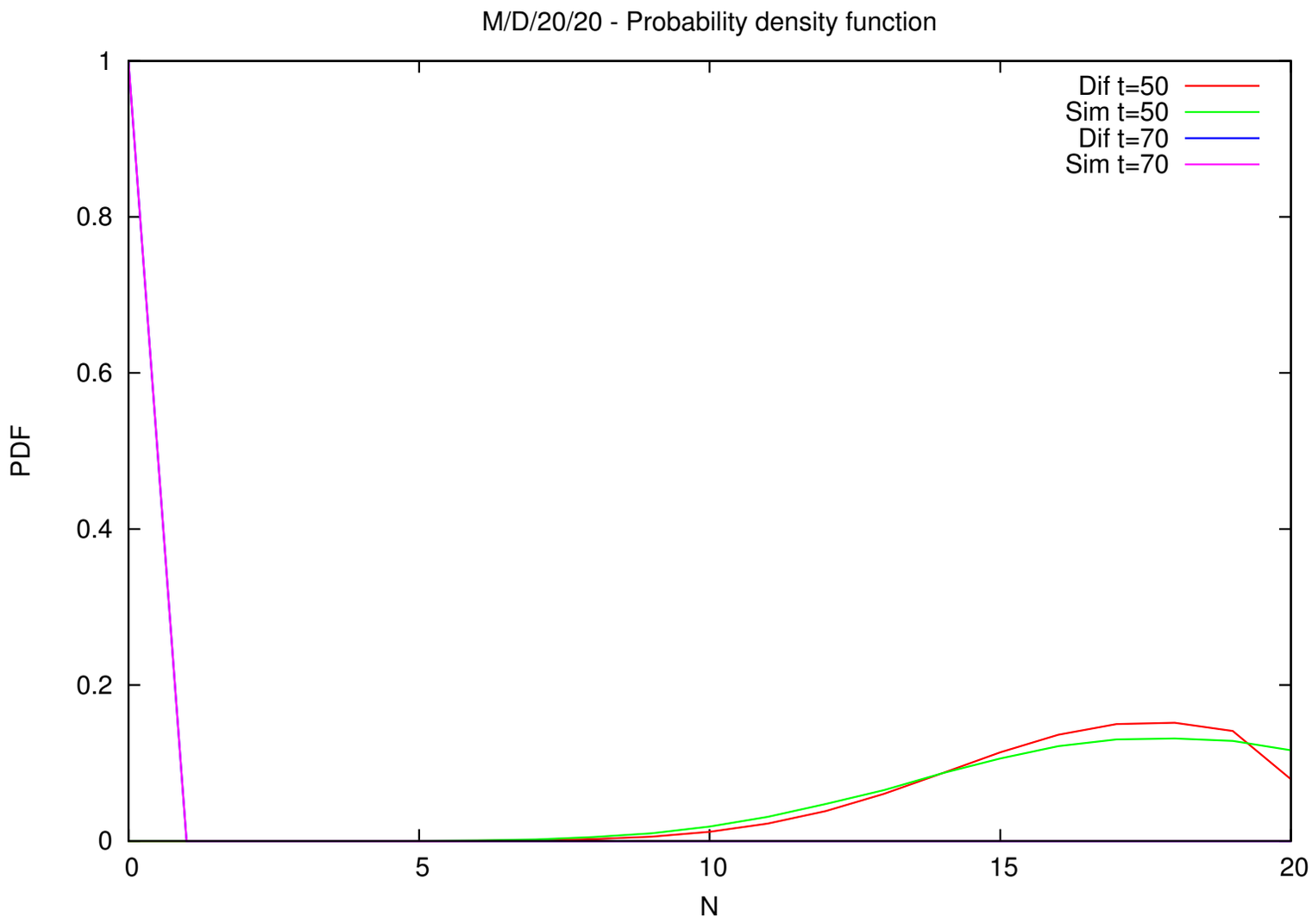


Figure 24: Densities $f(x, t = 50; 0)$, $f(x, t = 70; 0)$ and corresponding simulation histograms

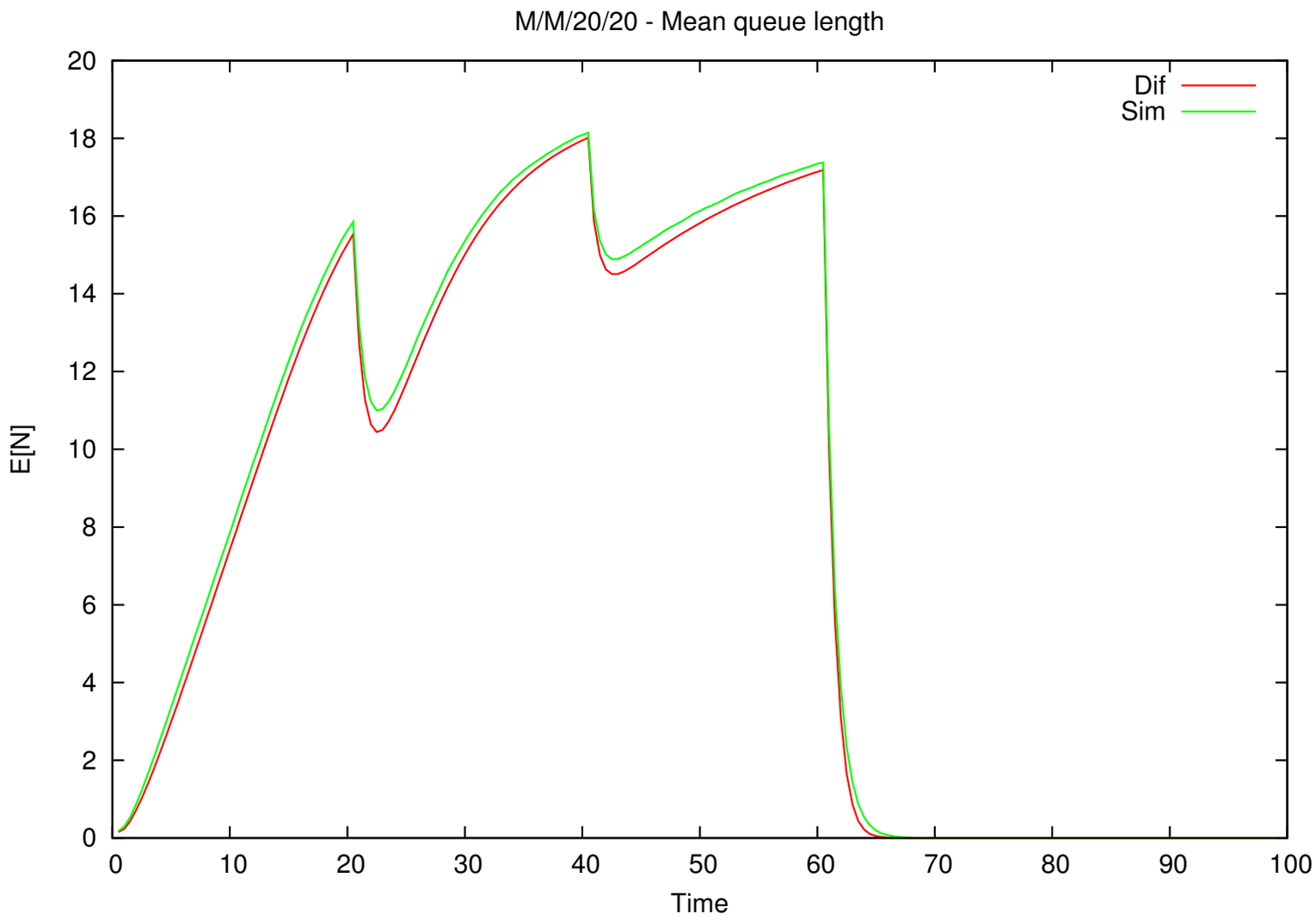


Figure 25: Mean number of customers as a function of time for M/M/20/20

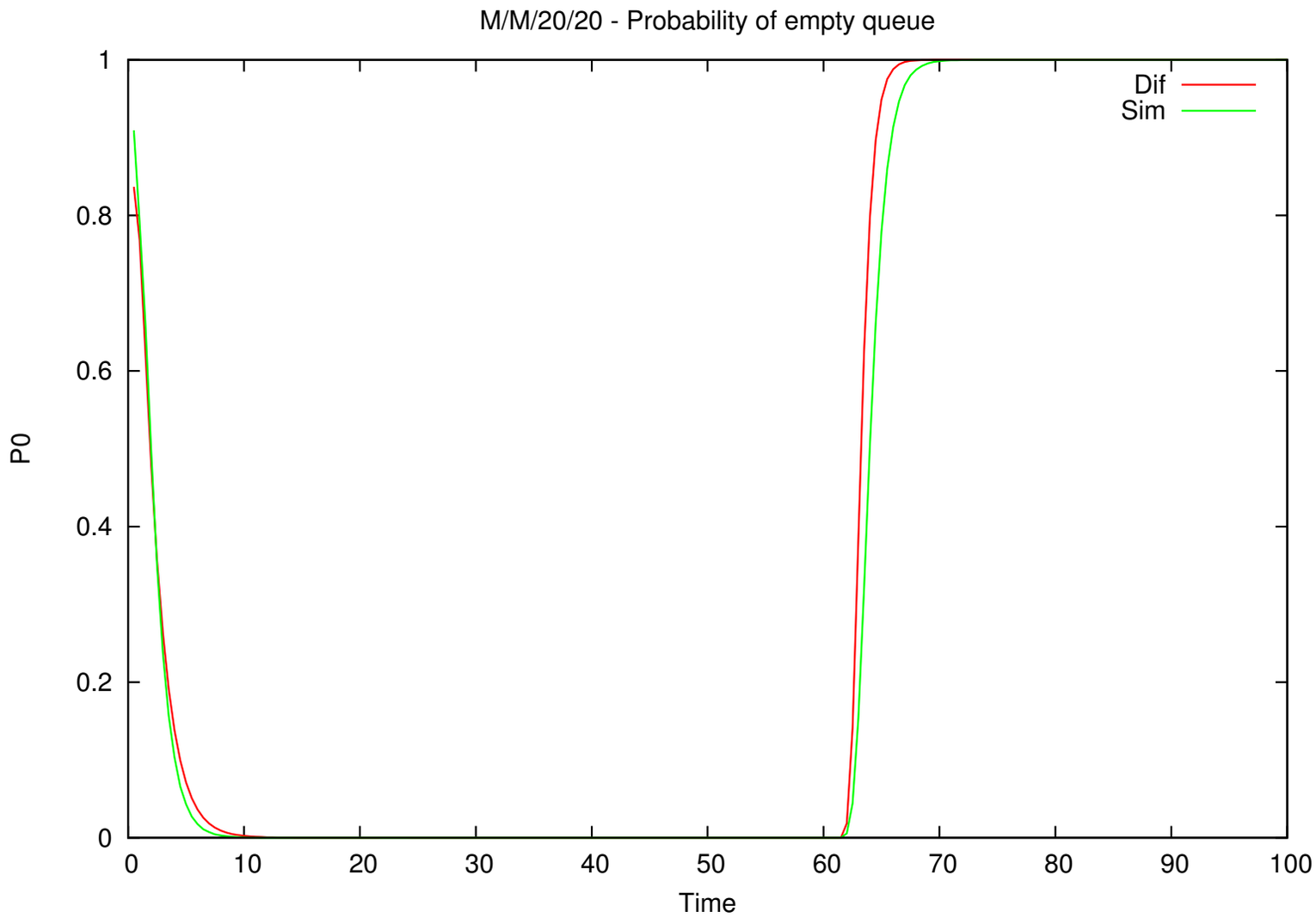


Figure 26: Probability $P_0(t)$ of an empty system for M/M/20/20

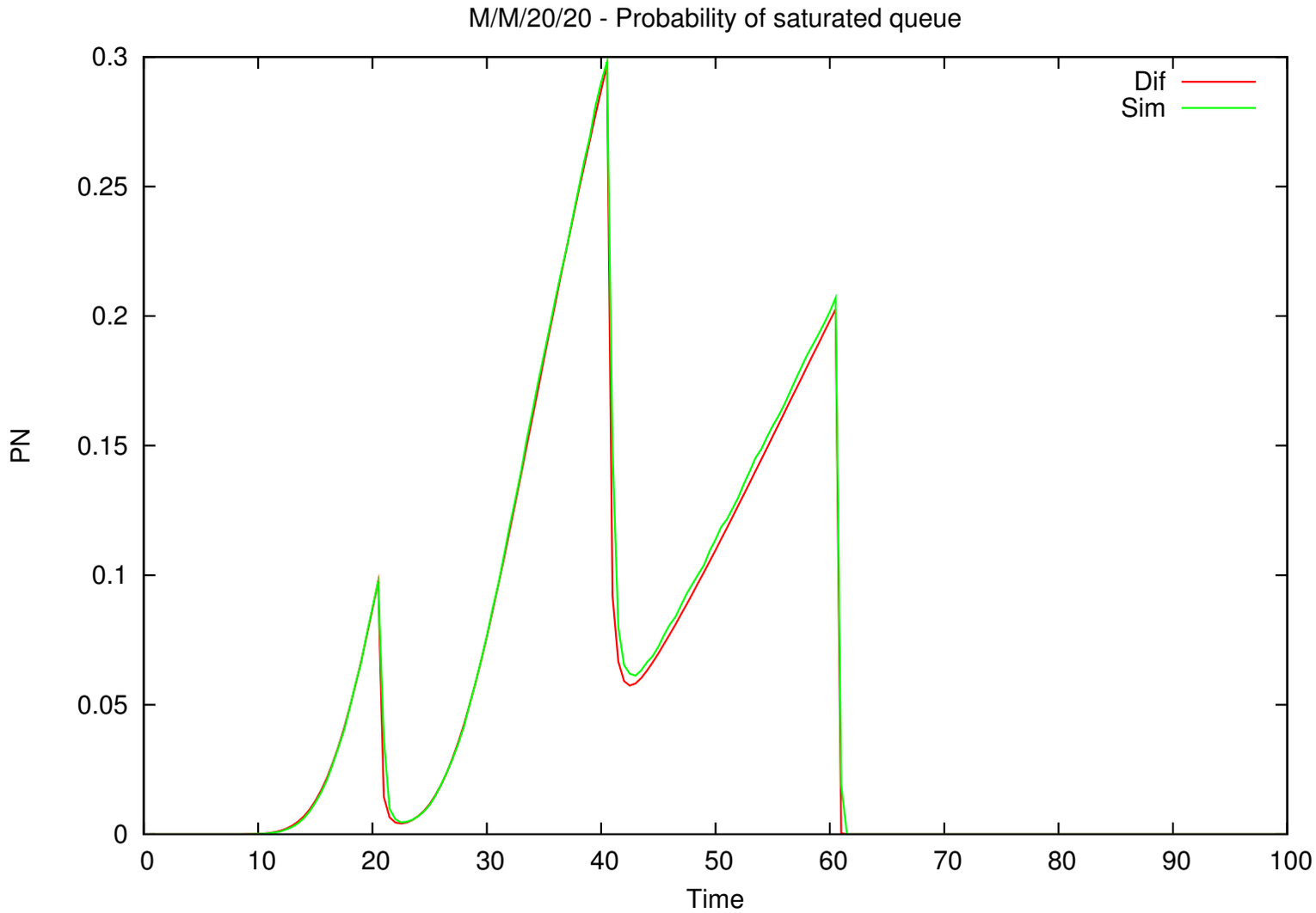


Figure 27: Probability $P_N(t)$ of a a full system (customer rejection) for M/M/20/20

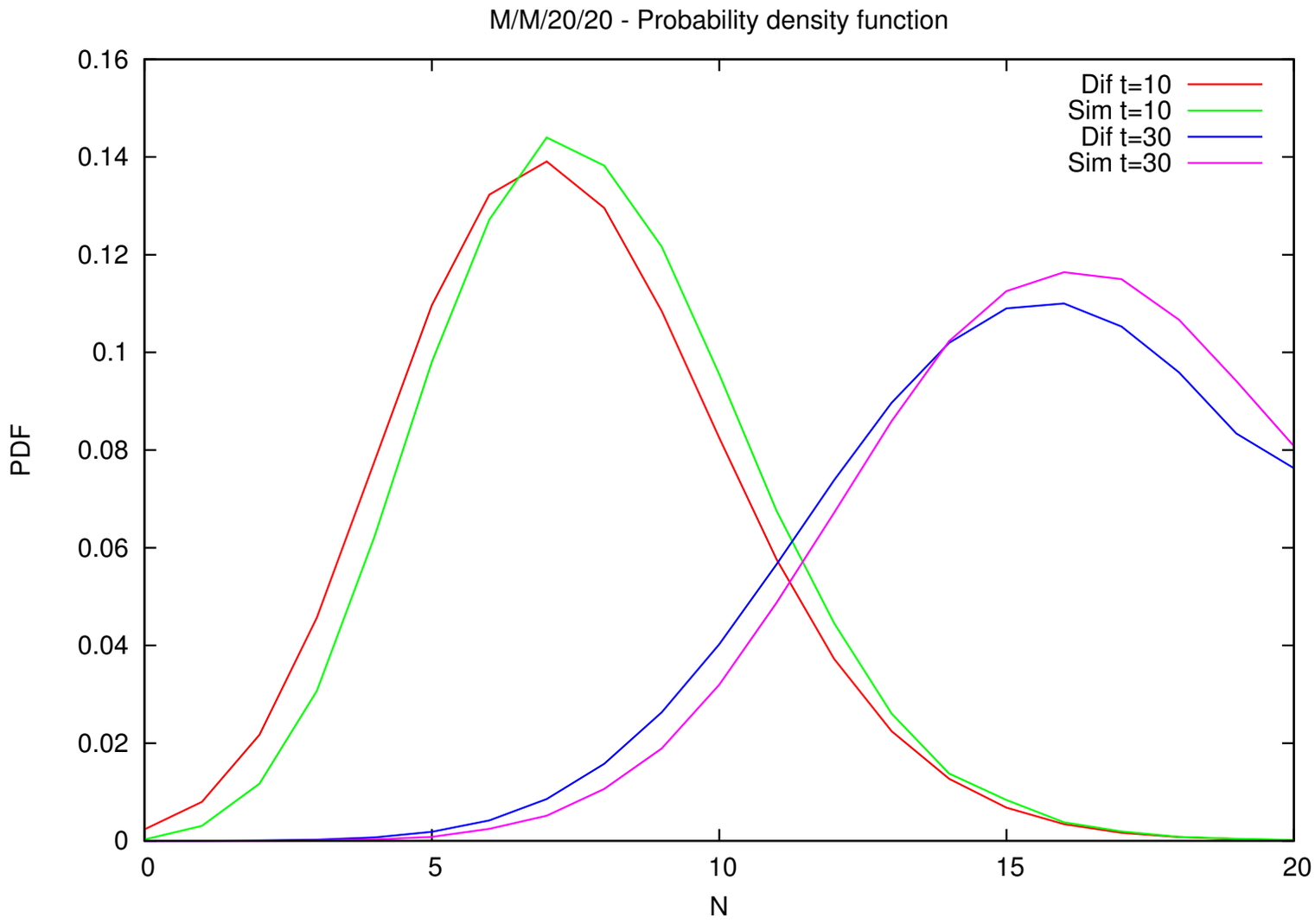


Figure 28: Densities $f(x, t = 10; 0)$, $f(x, t = 30; 0)$ and corresponding simulation histograms for M/M/20/20

Conclusions

- diffusion model of a single server assumes general interarrival and service time distributions this way going beyond Markov models,
- network models may have any topology, also hierarchical, and any number of nodes (are easy scalable),
- the results are obtained in form of queue distributions and waiting time distributions that makes easier to analyse QoS of paths, e.g. jitter,
- In a natural way it offers also transient state analysis and allows to predict the dynamic behaviour of the system under time-dependent load.
- the transient state model is solved step-by-step in small time intervals with parameters specific to these intervals; any decision of a controller concerning dynamic routing of packets, as well as changes of flows due to attacks and control mechanisms may be easily reflected in time-dependent and state-dependent diffusion parameters.